NORTH ATLANTIC TREATY ORGANIZATION

DEFENCE RESEARCH GROUP

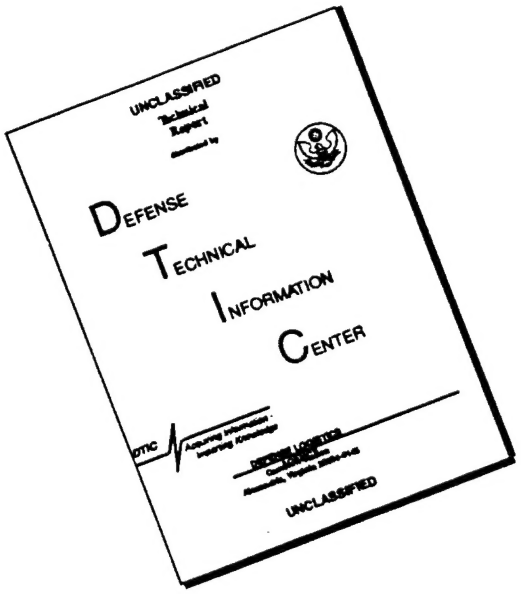
1095-96

TECHNICAL REPORT
AC/243(Panel 3)TR/21

POTENTIALS OF SPEECH AND LANGUAGE TECHNOLOGY SYSTEMS FOR MILITARY USE : AN APPLICATION AND TECHNOLOGY-ORIENTED SURVEY

Panel 3 on Physics and Electronics
Research Study Group 10 on Speech Processing

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

REPORT DOCUMENTATION PAGE

1. Recipient's Reference:	2. Further Reference:	
3. Originator's Reference: AC/243(Panel 3)TR/21	4. Security Classification: UU	
	5. Date: 14 August 96	6. Total Pages: 67

7. Title (NU):

POTENTIALS OF SPEECH AND LANGUAGE TECHNOLOGY SYSTEMS FOR MILITARY USE: AN APPLICATION AND TECHNOLOGY ORIENTED SURVEY:

8. Presented at:

9. Authors/Editors:

Dr. H.J.M. Steeneken, et al.

10. Author(s)/Editor(s) Address:

TNO Human Factors Research Inst
P.O. Box 23
3769 ZG Soesterberg
The Netherlands

11. NATO Staff Point of Contact:

Defence Research Section
NATO Headquarters
B-1110 Brussels
Belgium
(Not a Distribution Centre)

12. Distribution Statement:

Approved for public release. Distribution of this document is unlimited, and is not controlled by NATO policies or security regulations.

13. Keywords/Descriptors:

AC/243(PANEL 3/RSG.10), SPEECH TECHNOLOGY, SPEECH COMMUNICATION, SPEECH CODING, HUMAN INTERFACES, SPEECH RECOGNITION, TRANSLATION

14. Abstract:

The primary goal of this report is to describe the military applications of speech and language processing and the corresponding available technologies. The RSG.10 hopes that this report will be a useful tool for Operational staffs, Defence Research Staffs, and potential users within procurement departments of the NATO countries, by helping them to define and meet military requirements.
The military applications are itemized in six categories:
Command and Control, Communications, Computers and Information Access, Intelligence, Training which also includes language training, Joint Forces.
The description of available technologies includes:
Speech Processing, Language Processing, Interaction, Assessment and Evaluation,.
An overview of some case studies and applications is given.

CONSEIL DE L'ATLANTIQUE NORD
NORTH ATLANTIC COUNCIL

UNCLASSIFIED/UNLIMITED

ORIGINAL: ENGLISH
14 August 1996

TECHNICAL REPORT
AC/243(Panel 3)TR/21

DEFENCE RESEARCH GROUP

PANEL 3 ON PHYSICS AND ELECTRONICS

Technical Report on Potentials of Speech and Language Technology Systems for Military Use: An Application and Technology-Oriented Survey

1. This report presents the findings of Defence Research Group Panel 3 RSG.10 on Applications of speech Technology for Military Use. The military user and procurement communities are invited to review this report and its recommendations. Comments may be provided through the Defence Research Section at NATO Headquarters. The Preface and Executive Summary of this report are being circulated separately under reference AC/243-N/473 dated 23 August 1996.

(Signed) Dr. K. GARDNER
Head, Defence Research Section

NATO,
1110 Brussels.

This page has been left blank intentionally

Contents

Preface

Efficient speech communication is recognized as a critical and instrumental capability in many military applications such as command and control, aircraft and vehicle operations, military communication, translation, intelligence, and training. The NATO research study group on speech processing (AC243(Panel 3)RSG.10) conducts since its establishment in 1978 experiments and surveys focused on military applications of language processing.

Guided by its mandate, the RSG.10 initiated in the past the publication of overviews on potential applications of speech technology for military use (Beek et al. 1977, Weinstein 1991) and also organized several workshops and lecture series on military-relevant speech technology topics.

In recent years, the speech R&D community has developed or enhanced many technologies which can now be integrated into a wide-range of military applications and systems:

- Speech coding algorithms are used in very low bit-rate military voice communication systems. These state-of-the-art coding systems increase the resistance against jamming;
- Speech input and output systems can be used in control and command environments to substantially reduce the workload of operators. In many situations operators have busy eyes and hands, and must use other media such as speech to control functions and receive feedback messages;
- Large vocabulary speech recognition and speech understanding systems are useful as training aid and to prepare for missions;
- Speech processing techniques are available to identify talkers, languages, and keywords and can be integrated into military intelligence systems;
- Automatic training systems combining automatic speech recognition and synthesis technologies can be utilized to train personnel with minimum or no instructor participation (e.g. Air traffic controller).

In this report we review the wide range of potential military applications and we also describe the current state-of-the-art in speech technologies. The RSG.10 hope that this report will be a useful tool for Operational staffs, Defence Research Staffs, and potential users within procurement departments of the NATO countries, by helping them to define and meet military requirements.

This report is the result of the contributions of all RSG.10 members which represent nine NATO countries (Belgium, Canada, France, Germany, the Netherlands, Portugal, Spain, United Kingdom, and the United States). A list of the National Points of Contact and of the authors of this report is given in the appendix.

Because speech technologies are constantly improving and adapting to new requirements, it is the intention of the RSG.10 to update this document regularly to reflect this evolution. Therefore the RSG.10 appreciates any comments and feedback on this report.

This page has been left blank intentionally

Executive Summary

Summary and Major Conclusion of the Study

(i) The primary goal of this report is to describe the military applications of speech and language processing, and the corresponding available technologies. The military applications are itemized in six categories:

- Command and Control,
- Communications,
- Computers and Information Access,
- Intelligence,
- Training which also includes language training,
- Joint Forces.

For each category a description of the requirements and possible goals are given. The available technologies are subdivided in:

- Speech Processing,
- Language Processing,
- Interaction,
- Assessment and Evaluation.

For these technologies the state-of-the-art with respect to performance and availability is discussed. For speech processing a sub-division for speech coding, speech synthesis and recognition is made.
Also an overview is given of possible assessment procedures and design criteria. Finally some case studies and applications are described.

Major Recommendations

(ii) In brief the reports highlights the need of speech control for operational systems and advanced communications in a changing military environment. Reduction of personnel, increasing complexity of systems, multi-national operations require optimal human performance in which speech can be a natural means of interfacing.
The Research Study Group which performed this study hopes that it will be a useful tool for the Operational staffs, Defence Research Staffs, and potential users within procurement departments of the NATO countries.

Military Implications

(iii) Efficient speech communication is recognized as a critical and instrumental capability in many military applications such as command and control, aircraft and vehicle operations, military communication, translation, intelligence, and training. The study further provides some insight on the power and limitations of current speech technology and suggests areas of applications and potential areas where research in needed.

Future Work

(iv) The NATO research study group on speech processing will continue to carry out collaborative studies on the performance and application of speech technology in the adverse military environment. Specifically stress, high noise levels, vibration, g-forces, and non-native talkers are typical for the multi-national military environment.

Chapter 1. Introduction

Speech can be considered as the most convivial means of communication between humans. But the speech signal also carries information about the speaker (gender, identification), his or hers emotion, and the language spoken. It is therefore not surprising that speech technology embraces a wide range of applications in the military and civil world.

Defence applications introduces challenges that are not always met by commercially available systems. For example, automatic speech recognition used in military environments must be robust to adverse conditions. This is because many military situations involve difficult acoustic and mechanical environments such as high and variable noise, vibration and g-force levels. The RSG.10 has studied these military specific problems since its inception in 1978.

The primary goal of this report is to describe the military applications of language processing and the corresponding available technologies. The performance achieved by the various speech processing technologies is also described and examples of successful integrations in military systems are given.

The military applications are itemized in six categories in this paper:
- Command and Control;
- Communications;
- Computers and Information Access;
- Intelligence;
- Training which also includes language training;
- Joint Forces.

The available LRE technologies (Language Research and Engineering) include Spoken Language systems, Language processing, and Interaction between systems. A functional relation between the application and the corresponding technology required to apply speech or text as medium or control signal is provided by Table 1.1. A reference to the corresponding chapter and section is given as well. Table 1.1 also enables readers to quickly find the relevant information they are looking for. This table shows the relationship between military applications and the available technologies.
The specific areas of military applications, focused on the use of speech technology are described in chapter 2, the available technologies are described in chapter 3.

For a successful realization it is of great importance that the military requirements and specifications match with the performance offered by the systems. Adverse military conditions such as poor communications, noisy environments, vibration, stress, non-native speech, may deteriorate a speech signal or a text string. Assessment and evaluation considerations are described in chapter 4. Some available applications are described in chapter 5.

Table 1.1. Overview of areas of military applications of speech and language processing in the relation to available technologies. The numbers refer to the corresponding paragraphs.

Speech and Language in Military Applications \ Available Technologies	3.1 Speech Processing	3.1.1 Speech Coding	3.1.2 Speech Enhancement	3.1.3 Speech Synthesis	3.1.4 Speech Recognition	3.1.5 Speaker Recognition	3.1.6 Language Identification	3.2 Language Processing	3.2.1 Topic spotting	3.2.2 Translation	3.2.3 Understanding	3.3 Interaction	3.3.1 Interactive dialogue	3.3.2 Multi-model communication	3.3.3 3-D Sound Display
2.1 Command and Control		•	•	•		•				•			•		•
2.2 Communications		•	•							•					•
2.3 Computers and Information Access				•	•					•					
2.4 Intelligence		•	•			•	•		•	•			•		•
2.5 Training				•	•					•					
2.6 Joint Forces		•	•							•			•		•
5.0 Case studies															
5.1 Cockpit Fast jet		•	•	•	•	•							•	•	•
5.2 Helicopter		•	•	•	•	•							•	•	•
5.3 Sonar					•	•								•	•
5.4 Noise reduction		•	•												
5.5 Training of air traffic controllers				•	•								•		•
5.6 Spoken Language Systems demonstration				•	•		•			•	•		•		•
5.7 Voice Technology in Space				•	•				•				•	•	•
5.8 Speech Coders 600-1200 Bps		•	•												

Chapter 2. Speech and Language in Military Applications

The key military applications areas for speech and language technology, as indicated in Table 1.1, are: command and control; communications; computers and information access; intelligence, training; and joint (coalition) forces. In general, all the speech and language technology areas have some application to all the application areas, but Table 1.1 highlights particularly-important connections between applications and technologies. The purpose of this chapter is to briefly discuss each of the applications areas from the viewpoint of the requirements placed upon speech and language technologies.

In the multi-national NATO context, joint or coalition military operations make the need for multi-lingual speech and language technology particularly important. For example, speech recognizers must operate in the languages of all the forces. Also, in many situations, it will be necessary to process linguistic information (speech or text) in one language and provide information which is translated into another language or languages.

This need for multi-lingual operation adds to the usual military requirements for security, robustness against noise and jamming, and limited bandwidth channels. Also, speech communication systems must operate with high performance despite the fact that many of the talkers and listeners will be working in languages which are not their native language. This causes difficulty due to foreign accents and due to human comprehension limitations in non-native languages.

The following sections outline military applications and associated requirements in each of the six areas listed above. In summary:

- Command and Control can be aided by human interaction with computers, weapons and sensor systems by voice. But this application requires high performance of speech and language technology in real time, under adverse conditions including motion and noise, and with multi-lingual input and output.

- Communications must operate securely, with high intelligibility, under conditions of noise and jamming. The speech signal, for example, must be coded and transmitted with sufficient fidelity to be understood by listeners who are not native in the language being spoken.

- Computers and Information Access are a crucial part of modern military operations. Speech and language technology can be used to allow military personnel to command and query computers and information sources by voice, which will be particularly important for hands-busy, eyes-busy situations. Requirements on the technology include multi-lingual input and the possibility of translation or summarization of the information from one language into another.

- Intelligence places high demands on information processing, including processing of speech and text.

- Training of forces for military operations can be significantly aided by applying speech technology to allow people to interact with advanced simulation systems by voice. In addition, for joint (coalition) operations, training in foreign languages is essential; such training can be aided by utilizing speech and language technology to provide machine-aided foreign language exercise and tutoring for military personnel.

- Joint Forces operations require the coordination of forces speaking multiple languages. Here speech and language understanding and translation technology have great potential to increase the efficiency of operations. However, the demands on the technology are high, and initial applications probably need to focus on limited domains for translation and multilingual information exchange.

2.1 Command and Control

Command and Control is mainly concentrated around human operations. However, optimal functional behaviour can be supported by advanced interactive systems. A full automatic armament system is neither possible nor credible. One of the reasons for that is that they will never be as efficient as a human operator, in particular as far as adaptability under unknown situations, analysis on subtleties or details are concerned. Consequently the operator will have a major role in the general command and control loop: perception → processing → decision → action. This loop is presented in Fig. 2.1.

However, the fact of having an operator in the loop always implies limitations with regard to the handling of a situation or a system. This can be defined by both the physical dimensions and abilities of a human operator and by the limited capacity to process information and to respond to it. For example: the body of the operator is sometimes crucial for the design of a system. If an armoured car or cockpit is designed one has to take into account the fact that a person has to fit inside. For any system design one knows that a person has a fixed means of interaction (arms, fingers, legs, ears, eyes) sometimes impaired due to illness, stress or tiredness. Moreover, in order to be as efficient as possible, an operator has to be well trained, preferably in a very short time period.

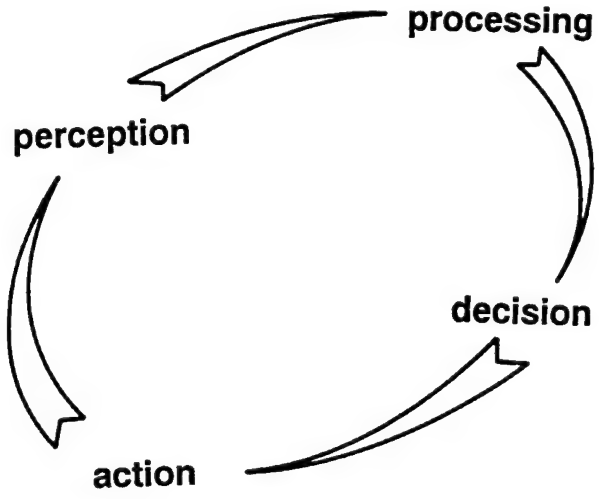


Fig. 2.1. Control loop of human decision making.

Although there is no generic solution to reduce the effect of these limitations, speech may help in some specific conditions: control of systems, reducing workload and training acceleration. The analysis of a generic armament system results in the definition of four top level functions:

- motion;
- communication;
- survival;
- specific function related to a particular system.

For each of these functions, speech technology has a potential area of application. The framework given in Table 2.1 considers two kinds of applications, that are representative of the various situations: C^3I-station and crews in a tank or an aircraft. Those two examples are described in terms of the four top level functions as mentioned before.

Table 2.1. Functional analysis of two command and control examples.

	C^3I station	Tank
Motion	C^3I station mainly stationary by itself	Principal functions: * mission planning and control * navigation * driving
Communication	The core function of C^3I: communication with the forces, communication with other C^3I stations	Three levels of communication: * internal, between the crew members * external, with other units or commands * control of on board systems
Survival	* pollution of information, implies secure data links * self destruction, no exploitation by the enemy	* mobilisation (see motion) * communication security and protection against jamming
Specific functions	potential ones (depending on the level the station is installed): * action planning * global situation presentation * intelligence collection	potential ones (depend on the mission assigned): * fire and destruction power * reconnaissance * fire excitation

Even if all the armament systems can benefit from the addition of tools based on speech technology, it does not imply that speech can be useful for each of the identified levels or functions. For example, it does not make any operational sense to drive a tank entirely by voice. On the contrary tracking tasks operated by voice are inferior in performance with respect to tracking by hand. Speech is not adapted for this, as it has not the ability for immediate feedback and a fine tuning of a parameter. However, speech can be very useful to control discrete and non critical functions, e.g. radio channel selection, check list control, data input and management, etc.

Each potential application has to be properly analysed in order to know if and to what extent speech can help, given the potentialities of speech technology. A thorough Human Factors analysis is therefore required sometimes driven by a Wizard of Oz experiment[1].

[1] An experiment in which the performance of a system is predicted by performing measurements in which crucial functions are simulated. For example: a system including automatic speech recognition can be simulated by using a human listener as recognizer.

Integrating speech into a system is not trivial. It has to be taken into account at the very beginning of the design of the global interaction between operator and system. The design considerations to be taken into account may include:

- Why use speech? What are the benefits of speech for the operator: improvement in accuracy, response time, workload decrease? In particular, the benefit of speech versus any other means of interaction has to be assessed.

- For what functions? Which interactions are to be mediated through speech, bearing in mind the impact of security, workload and usefulness (it is not necessary to build a function which is 100% speech-based if it is seldom used)?

- For which speakers? This implies to know whether one or several persons will use the system. Consequently recognition in either speaker dependent or speaker independent mode has to be chosen. It is also necessary to know the intellectual level of the users, since it has an influence on the choice of the means of interaction between the user and the system.

- Benefit versus costs: What will be the total cost for the system? This cost should include not only design and development costs but also operator training and maintenance costs.

First conclusions of various studies performed by the forces indicate that speech can be used for about 50% of the tasks in a C^3I station, but only for 25% in a vehicle (car but also aircraft). In particular, speech techniques are not yet robust enough to be used on security functions, where even a very low error rate can have dramatic consequences (for example if "eject" is understood in place of "reject" in an aircraft). Therefore, speech interaction can be envisaged for non critical tasks only. In this case, a recognition performance of 95% is in many situations satisfactory and *better* than human performance.

2.2 Communications

It is well known that the first speech transmission with electronic telecommunication means was realised in 1876 when Bell and Gray established the first telephone link on copper leads. Since this time the use of the telephone became obvious to prepare and to conduct military operations. A dialogue between two persons is more advanced than the exchange of written messages.
Presently we have wide band and narrow band communications. We can transfer written texts and complete files by means of electronic data links. However, personal communications with the possibility of a dialogue or immediate response is of crucial importance in many operational situations. Speech communication is spread all over a military organisation at all levels, but not all communications links are identical. The required level of security, the available radio-link, the possibility of being jammed are all important factors which define the system design. We can identify communications at staff level, and at tactical level:

a For interservice and headquarters communications the following features are identified:

- the possibility to establish long distance communications;
- a good intelligibility even if the speech signal is transmitted through heterogeneous transmission media (guided waves on coax-cable or on optical fibres, radio propagation above the ground, above the sea or in the air, free space propagation between earth stations and satellites);
- the possibility to reach heterogeneous terminal platforms (fixed headquarters; mobile headquarters; ships; planes);

- the use of conference and multimedia facilities;
- the transmission of secured speech by signal encryption;
- the use of translation facilities at both sides of the communication link.

All these features lead to the use of a network structure utilising several transmission means (telephone cable, optical fibre, line-of-sight links, satellite links). The speech signal has to be coded in a digital form to maintain a constant signal-to-noise ratio, independent of the length of the path. It is therefore mandatory to use the same coding technique at both ends of the communication link. This leads to the necessity to use common standards for speech coding. Such a standardization is obtained by the various NATO standardization agreements (STANAGs).
To reduce the bit rate, and the required bandwidth, future systems (especially those on satellite links) will use speech coding techniques with "vocoders", that are based on the production parameters of the vocal tract. These coding techniques could be combined with other services that also require an analysis of the speech signal. These new services are described in chapter 3.
New trends in the integrated networks consist in merging digital speech signals with other data signals coming from different kinds of information sources (data, video, fax, ...). It is now possible to realise a simultaneous communication of speech, graphics or image with multimedia terminals.

b Speech communication for intra-service or tactical use

In comparison with the users of the Tri-Service Staffs or of the Main Headquarters a user of speech communications in a tactical unit is interested in the following features:
- the use of mobile light-weight terminals;
- the possibility to reach simultaneously many sub units (broadcast transmission);
- acceptable quality of speech even in presence of electronic counter measures from an aggressor;
- the possibility to protect the spoken message by encryption.

These features lead to the use of radio nets that ensure the broadcast of messages. For military applications it is very important to send messages at the same time to different units. The use of the radio to transmit speech signals is very convenient for small units that have to move without constraints.

To protect these nets against the electronic warfare (EW) threats, spread spectrum communication systems, like frequency hopping radio nets, are used. These new techniques are not compatible with analogue signals. The speech signal has to be coded into a digital signal. To guarantee an interoperability with the inter-service communication it is mandatory to use the same coding scheme, and eventually the same encryption method, in all speech terminals. Advances in jamming technology have resulted in systems which are increasingly capable of disrupting and degrading critical communications. As a result, development and application of speech/audio technology for jam resistant communication systems is crucial in overcoming improved jamming capabilities. This technology is critical for effective battlefield communications and overall success. Future development of new speech compression techniques will provide ten to fifty times greater jamming resistance than current capabilities. Although the use of these techniques may constrain operations (e.g., limit the vocabulary), these compression capabilities will provide an effective countermeasure against jamming, as well as improve operation in the presence of self-jamming. Effective utilization of EW resources requires Electronic Warfare Support Measures (ESM) that are designed to determine the identification, location and disposition of opposing forces. Speech processing technologies such as speaker, language and keyword recognition can augment other technologies and improve the identification and tracking of opposing forces.

2.3 Computer and Information Access

All levels of military operations now require human interaction with computers and with data banks for entry or retrieval of information. Applications ranging from command and control, to logistics and maintenance, and to forward observer reporting, all will require computer access. Speech and language technology have great potential in these applications for allowing military personnel to query or command computers by voice.

The decreasing cost and increasing power of computer hardware and software have made the use of computers universal in the military. In addition, the quantity and range of types of available digital information have increased dramatically, to include not only text and numerical data, but also sound, graphics, images, and video. It is now possible to access and manage this kind of information with small, low-cost terminals. For example, the existence of digital terrain maps and compact presentation systems allow military aircraft pilots and navigators to make direct use of this type of information during their missions.

Now that we have more computers, more information, and more capability to access and process this information, the human/computer interface can become a bottleneck. Typically, people control access to the computer by keyboard, mouse, trackball, and touch panels. The output information is typically presented by displays or printers. But this type of input/output requires the use of the hands and eyes, which in military operations are often busy with other tasks. For example, it would be difficult for a forward observer to watch the objective and at the same time enter data via keyboard.

For such military applications, speech recognition can provide a useful input mechanism, while speech synthesis can provide a useful output mechanism.

Requirements on the technology for these applications include: multi-lingual speech recognition and synthesis, and high performance recognition under conditions of stress, workload, and noise. Specific applications include: repair and maintenance; control of auxiliary computerized systems; report entry for forward observers; and access to logistics data bases.

In addition to the human/computer interface, language technology can be very useful for converting information into a form which is understandable to the user. For example, machine translation technology provides the possibility for translating or summarizing information which is stored in a language foreign to the user, into the user's native language. Such multi-lingual information processing represents both a challenge and an opportunity for advanced language technology.

2.4 Intelligence

Intelligence requires processing of a large variety of types of information, including speech and text in numerous languages. With the growing complexity of the world situation, the number of languages in which information needs to be processed continues to increase. Also, the rapidly-changing nature of the world situation requires that information processing be able to adapt rapidly to new languages and domains. As indicated in Table 1.1, essentially all of the available speech and language technologies have potential application to this information processing activity.

2.5 Training

Well-trained personnel are imperative for the success of military missions. The definition of requirements for language and speech should take into account the following aspects:

2.5.1 Language Training

It must be differentiated between active and passive use of speech and language, e.g. in C^3I activities. New applications of speech and language processing are able to support training, using:
- computer-supported learning systems;
- production of computer-aided test facilities;
- production of didactic material.

2.5.2 Simulation

The need for simulated use of weapon systems stems from the immense costs involved in live operations. Therefore, speech and language applications must be integrated at all levels of simulated training. In particular, verbal man-machine dialogue, man-machine interfaces, voice-activated system command and feed-back will play a growing part in training activities.

2.5.3 Requirements

The following requirements emanate from the above-mentioned facts:
- reduction of training time and costs;
- increase of number of languages;
- improvement of proficiency levels;
- tele-learning;
- maintaining of language knowledge.

2.6 Joint Forces at Multinational Level

The change in the international military landscape results in a frequent use of multinational joint task forces.

The integration of units coming from several countries can cause a lot of language understanding problems at
- different moments of the coalition process:
- during the negotiation of the integration plans;
- during the preparations of training exercises;
- during the generation of mission orders;
- during the performance of the mission.

These problems arise due to the lack of:
- basic knowledge of the foreign languages involved;
- common definitions of critical military terms;
- employment of automatic machine translation systems.

Therefore, the players in the multinational joint forces theatre of operations need to be assisted by new means mainly based on speech and language technology, e.g.:
- Automatic translations of mission orders and messages to considerably reduce the reaction time of the military units;

- Conversion of voice reconnaissance data transmitted by a human observer into language-independent data;
- Automatic generation of specific military glossaries and multilingual dictionaries to be used at all staff and operational levels.

The major benefits of introducing speech and language technologies in a multinational military environment will be:

- enhancing the mutual understanding process;
- increasing the speed of operational exchange of multilingual information;
- reducing the reaction time of operational units in critical situations;
- speeding-up of foreign language acquisition and learning.

Chapter 3. Available Technologies

3.1 Speech Processing

Modern speech technology algorithms are based mainly on digital signal processing techniques. Currently near real-time processing by complex analysis methods, making use of digital signal processors (DSPs), allow us to perform feature extraction, statistical pattern matching and survey of large data bases. These techniques make it possible to perform significant data reduction for coding and transmission of speech signals. Also automatic recognition of speech, speaker or language can be performed. In this chapter the state-of-the-art is presented and related to realistic military applications.

3.1.1 Speech Coding

In the second half of this century, when digital systems became available, it was obvious that the transmission of digital signals was more efficient than the transmission of analogue signals. If analogue signals are transmitted under adverse conditions, it is not easy to reconstruct the received signal, because the possible signal values are not a priori known. For digital signals discrete levels are used. This allows, within certain limits, the reconstruction of distorted signals. The first digital transmission systems were based on coding the waveform of the speech signal. This results in bit rates between 8000 to 64000 Bps (bits per second). The higher the bit rate the better the quality. Later, more advanced coding systems were used where basic properties of the speech were determined and encoded, resulting in a more efficient coding (bit rates between 300 and 4800 Bps) but also in reduced intelligibility. These methods are discussed in this section.

The first technique used to convert an analogue signal into a digital signal was based on the work of Shannon. He converted the instantaneous signal value at discrete moments into a binary number (a sample) and proved that it was possible to reconstruct the original signal from these samples, if the sampling frequency is high enough. Theoretically the sampling frequency is required to be twice the highest frequency component of the analogue signal.

Based on this technique a conversion system was used for telephone speech signals (with a frequency range between 300 to 3400 Hz) by using a sampling frequency of 8 kHz. The conversion of the instantaneous signal value had a resolution of 256 discrete levels corresponding to 8 binary digits. These bits are then transmitted in series with a bit rate of 8×8000 Bps or 64000 Bps. This technique is known as Pulse Code Modulation (PCM), and is still in use. PCM is one of the methods used to realise time division multiplex (TDM) where bit streams of different channels are combined in order to transmit many simultaneous telephone links using the same transmission channel.

In order to reduce the bit rate, and thereby increase the number of simultaneous channels in a given bandwidth, it is necessary to increase the coding efficiency. Therefore, the signal is compressed before encoding at the transmission side, and expanded after decoding at the receiving end. There are presently two different compression algorithms in common use: the so-called A-law used in Europe, and the μ-law used in North America. The differences between the two methods are small and it is possible without much distortion to use one of the two methods at one side and the other method at the other side.

Another method to convert analogue speech signals consists of using a delta-modulator. In this case, the sampling frequency is much higher than twice the highest frequency component and only one bit of information is transmitted per sample, corresponding to the slope of the signal (differential quotient). By making use of a simple integrator, the original waveform can be

retrieved. This technique results in good signal quality at lower bit rates than is required for PCM. In general a bit rate of 16 or 32 kBps is used.

Further enhancements to these methods, including dynamic optimization, have resulted in the CVSD (Continuous Variable Slope Delta modulation) and ADPCM (Adaptive Differential Pulse Code Modulation) methods.

The relation between the instantaneous analogue value of the waveform and the digital representation is different for PCM and Delta modulation. For PCM, the most significant bit of the digital representation represents the biggest portion of the analogue value, hence digital errors are more dramatic if this value is distorted. For Delta modulation, all bits have an equal significance, making this method more robust to channel errors. PCM error rates of 1% will give an unacceptable deterioration, while for delta modulation error rates up to 15% will result in an acceptable quality.

Whereas waveform coders like the ones described above aim at a faithful reproduction of the signal waveform, vocoders explore our knowledge of speech production, attempting to represent the signal spectral envelope in terms of a small number of slowly varying parameters. Vocoders achieve considerable bit rate savings, being aimed at bit rates below 2400 Bps, however, they result in degradation of the speech quality and of the ability to identify the speaker.

Many new coding schemes have been proposed recently which could hardly be classified according to the waveform-coder/vocoder distinction. This new generation of coders overcame the limitations of the dual-source excitation model typically adopted by vocoders. Complex prediction techniques were adopted, the masking properties of the human ear were exploited, and it became technologically feasible to quantize parameters in blocks (VQ, vector quantization), instead of individually, and use computationally complex analysis-by-synthesis procedures. CELP (Code Excited Linear Prediction), multi-pulse, and regular-pulse excitation methods are some of the most well-known "new generation" coders in the time-domain, whereas in the frequency-domain one should mention sinusoidal/harmonic and multi-band excited coders. Variants of these coders have been standardized for transmission at bit rates ranging from 12 down to 4.8 kBps, and special standards have also been derived for low-delay applications (LD-CELP).

Today, audio quality which can be achieved with the so-called telephone bandwidth (3.4 kHz) is no longer satisfactory for a wide range of new applications demanding wide-band speech or audio coding. At these band-width (5−20 kHz), waveform coding techniques such as sub-band coding and adaptive transform coding, have been traditionally adopted for high bit rate transmission. The need for 8-to-64 kBps coding is pushing the use of techniques such as linear prediction for these higher band-width, despite the fact that they are typical developed for telephone speech. The demand for lower bit rates for telephone bandwidth is, however, far from exhausted. New directions are being pursued to cope with the needs of the rapidly evolving digital telecommunication networks. Promising results have been obtained with approaches based, for instance, on articulatory representations, segmental time-frequency models, sophisticated auditory processing, models of the uncertainty in the estimation of speech parameters, etc. The current efforts to integrate source and channel coding are also worthy to mention.

Although the main use of speech coding so far has been transmission, speech encoding procedures based on Huffman coding of prediction residuals have lately become quite popular for the storage of large speech corpora. For a recent exhaustive survey of speech coding, see Spanias (1994), which includes over 300 references to the relevant literature and comparisons of different schemes from 64 kBps down to 800 Bps, on the basis of MOS and DRT scores (see section 4.2) and computational complexity.

Military Applications

The applications of speech coders in military applications are driven by two concerns, security and bandwidth saving. The bandwidth considerations (and to some extent security) are driving similar interests in the commercial area.

The STU-3 (Secure Telephone Unit) provides a basic secure voice capability for telephone use. The original STU-3 units were equipped with a 2400 Bps LPC-10 vocoder. Newer units also include a 4800 Bps CELP coder. In addition manufacturers are allowed to include their own proprietary algorithms if they choose. During the secure call setup process, the two units negotiate automatically for the best algorithm supported at both ends and within the bandwidth limitation of the current telephone connection. Although the STU-3 is considered to be a useful, it is agreed that there is still a need for improved quality at 4800 Bps and especially at 2400 Bps.

Wireless communication is further driving the development of speech coding algorithms. The same needs are found both in tactical military and new commercial systems. The requirement to go from a hand-held terminal directly to a satellite imposes power and antenna size considerations which demand low bit-rate communication. Bit rates in the range of 2400−4800 Bps are typical of these requirements. Since these systems will often be used in mobile or other noisy environments, the algorithms must be robust to acoustic background noise and operate over fading communications channels.

There are other application areas which need further study:

- Applications involving anti-jam or low probability of intercept would benefit from coders operating at even lower rates. Coders operating at rates down to 300 Bps or even lower need to be considered.
- Conference among users using speech coders is an important requirement for command and control. Methods of integrating this capability into communication systems are the subject of current research.

3.1.2 Speech Enhancement

3.1.2.1 Communications in Adverse Conditions

Environmental noise, noise introduced by the transmission channel, band-pass limiting, non linear distortion, distortions in the time domain (echoes, reverberation), and bit errors deteriorate the speech signal and consequently the intelligibility. In military applications environmental noise and jamming are the major sources of disturbances. Also variation of the vocal effort (speech level) may affect the efficiency of a speech transmission system. In such a situation automatic gain control may be applied. If, however, the speaker is placed in a noisy environment (tank, aircraft) the signal-to-noise ratio at the input can also be improved by using an advanced noise cancelling microphone. In some situations a speech enhancement unit (see sections 3.1.2.2 and 5.4) offers an improvement in the signal-to-noise ratio (SNR). However, in many cases, enhancement of the signal-to-noise ratio can also deteriorate the speech signal.

On the output side the listener may also be situated in a high noise environment. By using a headset with a high attenuation of the environmental noise the speech transmission quality at the ear of the listener can be improved. Present technologies offer headsets with an active noise reduction system (adding the same noise in anti-phase). Such a system offers an additional sound attenuation up to 25 dB for low frequencies (diesel engines).

For a specific application (noise level and spectrum) the combination of transmission system and the electro-acoustic transducers (microphones and headsets) has to be evaluated. Assessment methods for this evaluation are given in section 4.2. In order to show the importance of the selection of a microphone and the optimal position near the mouth the following example of such an evaluation is given.

Microphone Performance in a Noise Environment

Noise-cancelling (gradient) microphones were developed for use in a high noise environment. The specifications, given by the manufacturers, normally describe the effect of the noise reduction in general terms and are not related to intelligibility, microphone position or type of background noise. In Fig. 3.1 the transmission quality, expressed by STI (Speech Transmission Index, an objective measure for intelligibility, see section 4.2.2), is given for two types of microphones as a function of the environmental noise level. For these measurements an artificial head was used to transmit the STI-test signal. The microphone was placed in front of this artificial head at a representative distance from the mouth. The test signal level was adjusted according to the nominal speech level. The head was placed in a diffuse noise field with an adjustable level.

From the figure we can see that the distance from the mouth is an important parameter. For good communication quality, an STI of 0.6 is required. The lower limit is 0.35. It is also obvious that the two noise-cancelling microphones have different performance when operating in the type of noise environment used in this example.

In addition to a good communication system, discipline from the user is required. Communication in adverse conditions requires proper positioning of the microphone and headset as well as preventing the simultaneous use of the press-to-talk switch by multiple users, which increases the noise level introduced by the other microphones.

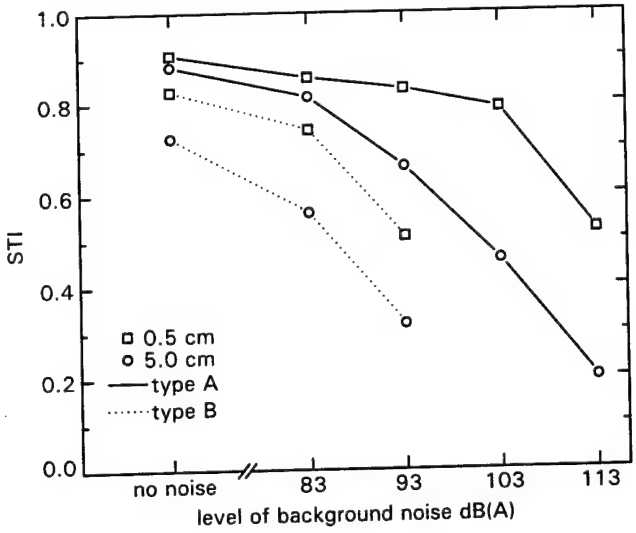


Fig. 3.1. STI as a function of the noise level for two different microphones and two speaking distances.

3.1.2.2 Enhancement Techniques

Many voice communication channels cause degradation of the speech signal. This is particularly true of military communication systems where noise interference is a common problem. When the speech signal is degraded, a loss in the intelligibility of the message and a reduction in the perceived quality of the speech signal usually occur. Speech intelligibility and quality improvements are therefore the two main objectives targeted by speech enhancement technologies designed for human listeners. Automatic speech recognition systems (ASR) suffer even more than human listeners from even slight degradations of the speech signal. A recent RSG.10 study has shown (Gagnon and Cupples 1995) that the utilization of speech enhancement technology such as noise reduction can dramatically improve the performance of ASR in noise as encountered in military environments.

For over 30 years, researchers have explored techniques to improve the quality and intelligibility of speech signals transmitted in the presence of interfering noise and channel distortions. The current availability of powerful digital signal processors allow the implementation of this useful technology in military communication systems. Speech enhancement technology can be used to increase the intelligibility and quality of speech signals transmitted over telephone networks, radio channels, and by speech coders. They can also be used to increase the intelligibility of noisy recordings obtained in intelligence gathering activities. They are also effective as pre-processors to speech and speaker recognition systems to extend the operational range.
As there is a broad variety of applications for speech enhancement and many different sorts of degradations of the speech signal, various approaches are exploited in speech enhancement systems originating in different types of speech enhancement technologies. Most existing practical systems integrate a combination of speech enhancement technologies as required by the application at hand.

Wideband Noise Reduction

A common form of degradation affecting speech signals is uninterrupted wideband noise. It is also one of the most difficult problems speech enhancement systems have to confront because this type of noise overlaps the speech signal in both the time and frequency domains. State-of-the-art wideband noise reduction technologies attempt to remove the noise using filtering techniques exploiting statistical models for both the noise and speech signals. Multi-state models enable noise reduction systems to deal with the non-stationary behaviour of both noise and speech signals (Gagnon 1993). Because wideband noise reduction systems increase the SNR of signals, they can dramatically increase the performance of ASR used in noisy conditions. For stationary white noise, the increase in performance is equivalent to a reduction of the noise by a factor of 6 to 12 dB. For human listeners this SNR increase results in a better perceived quality of the speech signal (better listenability). Unfortunately, it does not always materialize in better speech intelligibility. The ability to increase the intelligibility of speech signals is highly dependent on the noise characteristics and the inherent intelligibility that the speech signal carried before it was corrupted by noise. Current systems are effective for positive SNR values.

Impulsive Noise Reduction

Impulsive noise can be reduced efficiently by time-domain techniques or LPC based methods (Vaseghi and Rayner 1990). These systems identify the noise pulse and then replace the corrupted signal segment by estimating the signal based on neighbouring speech information. Some LPC based systems detect impulsive interference at the residual level which provides a 10 dB advantage in recognizing pulses which sometimes have smaller amplitudes than the corrupted speech signal. These systems are efficient if the interference pulse has a short duration and reasonable duty cycle ($<20\%$) compared to the speech signal.

Narrowband Noise Reduction

Narrowband noise (tonal noise) is usually removed using time-varying narrowband filters which identify tone interferences and filters them out. Narrowband noise is successfully removed when the interference does not overlap the formants regions and when the duration of the interfering tones is longer than the average speech phones durations.

Time-Frequency Modifications

In information gathering activities, military audio-analysts sometimes have to transcribe audio recordings where the speech intelligibility is degraded. Time-scale modification of the speech signal allow the analyst to playback the recording at different speeds without altering the pitch (perceived tone) of the signal. Slowing down a low intelligibility segment can help to enhance critical acoustic cues of the speech signal, resulting in a better understanding of the overall message. Inversely, speeding up during playback allows analysts to quickly find the important information in a recording. Such systems are continuously made better by the military research community. State-of-the-art systems can be used to change the playback rate of signals from half ($0.5\times$) to twice ($2.0\times$) the original speed without adding too many artifacts.

Many practical implementations are variations of time-domain methods which use inventive cut and paste techniques. Newer technology use short-time Fourier analysis or other frequency-time representations to modify the signal. There is a trade-off between the quality of the signal obtained and the computational load. Real-time implementations are possible for most algorithms using DSP technologies. Time-frequency modification can also be used in tandem with noise reduction to process low intelligibility recordings.

Channel Distortions

Channel distortions also reduce the performance of ASR systems. Many technologies exist to automatically compensate for channel distortions. Many algorithms used to render ASR systems more robust to linear channel distortions are variations of the technique known as blind deconvolution. Graphic equalizer can also be embedded in systems to compensate for channel distortions effects in human listener applications.

Speaker Separation

Co-channel interference suppression is an extremely difficult problem. Even in the ideal case of speech in the quiet, it has been difficult to demonstrate improvement in formal intelligibility tests.
Most techniques attempt to do separation by exploiting some sort of pitch synchronous analysis to separate voiced speech segments when speech from one or both speakers is voiced. Once speech is separated on this basis, it is necessary to determine which segments belong to which speaker. Zissman applied speaker identification techniques to show that one can determine whether one or both speakers is talking at any instant, given models of the individual speakers (Zissman 1991). This may or may not be a reasonable assumption in a practical system. These speaker identification techniques can also be used to identify which pieces of the separated speech belong to which speaker. One system claims slight improvement in word recognition based on tests with human listeners (Naylor and Porter 1991).
To illustrate how these technologies can be integrated in a real-time system see section 5.4 in which an interactive noise reduction system, GRYPHON, is described.

3.1.3 Speech Synthesis

In the area of speech synthesis, it is important to differentiate between two main categories: the first category is generally known as "canned speech", because the output speech is generated on the basis of pre-stored messages. The use of coding techniques to compress the message is also fairly common in order to save storage space. With this type of synthesis, very high quality speech can be obtained, especially for quick response applications. This technique, however, requires the use of large memory and is not very flexible. The second category, "text-to-speech synthesis", allows the generation of any message from text. This generally involves a first stage of linguistic processing, in which the text input string is converted into an internal representation of phoneme strings together with prosodic markers, and a second stage of sound generation on the basis of this internal representation. The sound generation can be made either entirely by rule, typically using complex models of the speech production mechanism (formant synthesis), or by concatenating short pre-stored units (concatenative synthesis). The speech quality obtained with concatenative synthesis is generally considered higher, however, the memory requirements are larger than with formant synthesis. Although the development of a new "voice" for the concatenative synthesizer can be to a large extent automated, concatenative synthesis is also less flexible in terms of changing speaker characteristics.

The current technology in text-to-speech synthesis already yields quality that is close to human speech, unlike older systems, which were unnatural sounding. Synthetic speech, although intelligible, is still clearly inferior to human speech in terms of naturalness and the expression of emotion.
Much research is required in these areas, however, the assessment of text-to-speech systems is very difficult because of the subjectivity of human listeners. Tests should be designed to determine the intelligibility and naturalness of the synthesized speech.
Text-to-speech systems with good quality are available for many languages and require moderate computer resources (a normal PC equipped with an audio board may be sufficient).

3.1.4 Speech Recognition

Automatic speech recognition (ASR) is the capability of a machine to convert spoken language to recognized words. These words are processed by an understanding stage and results into one or more actions. The action, which is a function of the application, could be for example the tuning of a radio receiver, a request for information or the conversion of a spoken input to text. Whatever the action, ASR can be valuable where the user's hand and/or eyes are busy performing other tasks, such as a pilot flying an aircraft.

3.1.4.1 Speaker Dependent or Speaker Independent Recognition

Speech recognizers can be speaker dependent or independent. Speaker dependent recognizers are trained by the user (person who speaks to the recognizer). Although the recognition accuracy is better than it is for speaker independent recognizers, the user is burdened with the training process (that is, providing multiple spoken examples of speech units, words and/or phrases needed to perform the desired action). Training may be particularly difficult for the user when the vocabulary becomes hundreds of words. Speaker dependent recognizers that automatically adapt to the speaker characteristics reduce the training requirements during use. Such systems are often called speaker adaptive systems and are delivered to the user with a factory trained vocabulary.

Speaker independent recognizers attempt to maintain recognition accuracy independent of who speak the words. An advantage of speaker independent recognition is that no training by the user is required. A disadvantage is that recognition accuracy is generally less than it is for speaker dependent recognizers for the same vocabulary size.

3.1.4.2 Isolated and Connected Speech Recognition

The difference between isolated and connected speech is that in the first case the speaker must leave distinct pauses between each word. This forces the user to speak unnaturally, but can be useful in many applications, such as command and control or under high noise conditions. In the second case the user speaks naturally with normal conversational or breath pauses. Connected speech makes the recognition much more complicated than for isolated speech because it is more difficult for the system to detect the word boundaries. This may result in reduced performance relative to an isolated recognition system.

3.1.4.3 Vocabulary Size

Classically, there has been a distinction in automatic speech recognition between "small vocabulary" and "large vocabulary" systems. The vocabulary of an automatic speech recognition system is defined as the set of words that the systems is capable of recognizing. For the purpose of this paper we will define a small vocabulary ASR as a system for which all words in the vocabulary must be trained at least once. Large vocabulary systems recognize sounds (characteristic parts in the speech signal) rather than whole words. It has a dictionary which contains the pronunciation of all words in the vocabulary. Large vocabulary systems are capable of recognizing words that have never been in the training set.

In large vocabulary systems the recognition process starts by recognizing the speech sounds in the input signal and the sequences of sounds are then recognized as words. The latter step is generally carried out by looking up the pronunciation of words in a pronunciation dictionary, also known as the lexicon. This lexicon must have at least one entry for each of the words in the vocabulary. Assembling a lexicon is a laborious effort, and must be performed very carefully. The performance of the ASR depends strongly on the quality of this dictionary.

The most direct application of a large vocabulary ASR is that of dictation. For this purpose, the vocabulary must be large enough to cover most of the words that are going to be used in the dictation task. If the domain is not exactly known, then vocabulary must cover almost an entire language. For English, a lexicon of 20,000 of the most common words provides coverage on the order of 97.5% of the words used in newspaper texts. The comparable coverage for French is 94.6%, Dutch 92.8% and German 90%. For certain domains, such as the military, the number of words needed in the vocabulary is generally less than 1000 words, due to formal or procedural use of language.

If all possible speech sounds in a large vocabulary ASR have to be trained for the individual phones (typically 35-50, depending on the language), this training can become very tedious for an individual user. Therefore ASR systems can come with pre-trained models, that can be adapted to fit the specific features of a certain user.

3.1.4.4 Phonetic Feature Based Recognition

Phonetic feature-based extraction is an optional stage in some speech recognition systems. In these feature-based systems, speech is modelled as a phoneme string, each phoneme being described by a set of phonetic features. The segments of acoustic signal corresponding to successive phonemes

can more or less overlap. The relations between the various phonetic features and the observed acoustic signal are stochastic and the corresponding statistics can depend on the context (coarticulation, stress, ...). Several benefits can be expected from such a feature-based model:

- Reduction of dimensionality: each part of the speech signal corresponding to a phonetic feature (the so-called acoustic cues) has a dimension far lower than that of the whole signal and consequently the size of the training set can be far lower too.

- The coarticulatory effects can be taken into account in a more elegant and concise way than just considering as many triphones as necessary. This advantage could be extended to other causes of speech variability.

- It must be underlined that the feature-based approach could decrease the computation load needed for lexical access by using broad phonetic features in a first step and the remaining features only when necessary.

3.1.4.5 Large Vocabulary Speaker Independent Connected Speech Recognition

Large vocabulary connected speech recognition systems often are speaker independent. Because of the large variability in speech produced by different speakers, these ASR systems have to be trained with many examples of speech, uttered by many different speakers. The state-of-the-art training databases for US English now consist of 66 hours of speech spoken by 284 different speakers.

When the recognizer has found a series of speech sounds, this sequence of speech sounds can correspond to many combinations of words that fit the sequence reasonably well (some deviations have to be allowed because the phones may not have been recognized perfectly). To humans, most of these word combinations do not make any sense and in order to disambiguate these possible combinations the recognizer is equipped with a language model. Language modelling is a separate topic of linguistic engineering and involves statistical analysis of the ordered occurrence of words in the domain.

In the same way that the speech sounds were trained, training of the language model is performed by counting word combinations in a sample of a training text. A larger amount for training text generally results into a more accurate language model. Currently, the largest (English) language models are trained on newspaper text covering several years, comprising a total of 250 million words. But even with that many words, many possible combinations of words have not occurred, so that estimation of those probabilities must be made using other techniques.

Most of the development in large vocabulary, connected speech recognition has been made in (American) English. Apart from a large potential market for dictation systems, the US organization ARPA (Advanced Research Projects Agency) has played a stimulating role in the development. ARPA not only sponsored speech recognition laboratories in their development of new systems and algorithms, they also organize a competitive assessment on a yearly basis, in which the performance of the various ASR systems is compared by an independent organization. ARPA sponsors the very labour-intensive task of recording acoustic training data bases, and makes sure that all training material used in the benchmark test are available for all competitors. These yearly assessment periods, together with all efforts in data collection, has had a tremendous influence on the performance of the ASR systems.

In Fig. 3.2 the performances of the best systems are given. The results were obtained in the various ARPA benchmark tests for speaker independent, continuous speech recognition systems. In

1987, the best performing system misrecognized 1 out of every 9 words, on the average, in a "Resource Management" task; a Navy application with a 1000 word vocabulary. The language model consisted of a simple word-pair grammar [the "probability" that a certain word can follow another is either 0 (impossible) or constant (possible)]. Each year the word error rate dropped, and in 1992 it was decided that a more difficult task was necessary. This became a dictation task with two conditions: a 5000 word vocabulary and an "open" or "unlimited" vocabulary. The test -sentences were read from the Wall Street Journal and other North American Business papers. The most recent data point for the hardest task is a 7.2% word error rate, or about 1 on every 14 words incorrect.

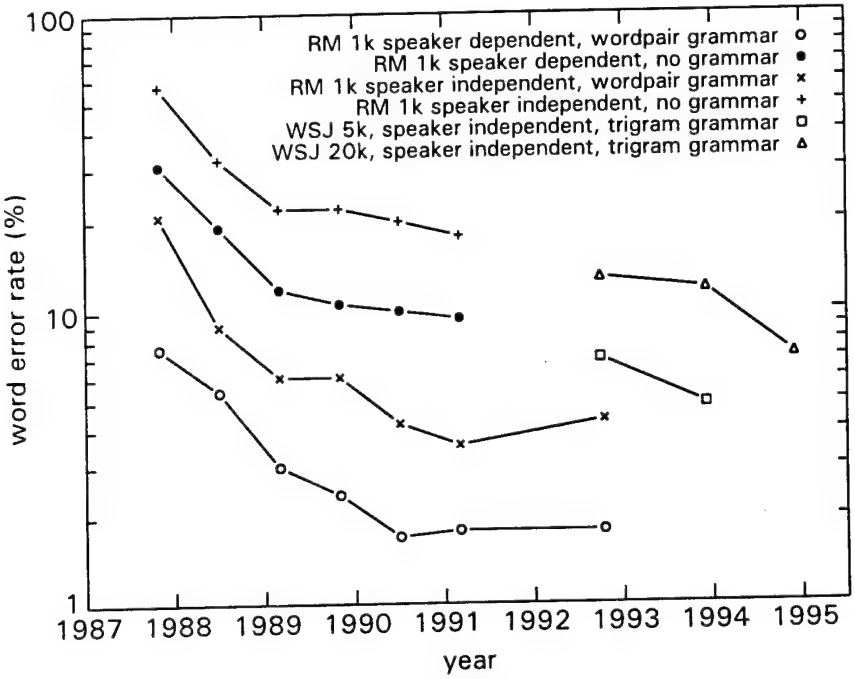


Fig. 3.2. Performance of state-of-the-art speech recognizers as a function of the year of evaluation obtained from the ARPA benchmark tests.

It is interesting to note, that recognition performance is comparable to human performance. In a pilot experiment conducted in the European project SQALE (Van Leeuwen et al. 1995), human listeners were offered sentences very similar to the ARPA tests, and they were asked to type in what they heard. For non-native English speakers (Dutch students) the average word error rate was 7%, while for natives (British and US students) this was 2.6%.

Although the current state-of-the-art is very impressive, there are still many practical problems to overcome. First, these results were obtained under noise-free conditions, with read speech (as opposed to more natural spontaneous speech). Under adverse conditions, such as stress, cross-talk, "incorrect" use of grammar, the word error rates increase. Second, the use of the algorithms in real-time require computational power not yet obtainable in a practical package. A trade-off can be made by accepting lower word accuracy to allow real-time operation in a useable package.

3.1.4.6 Robust Recognition for Adverse Conditions

The speech signal is deteriorated in many situations by factors which impact on the quality of the speech produced by the speaker or by the transmission equipment or channel between the speaker

and the recognition system. Some well known distortions introduced by the speaker are the Lombard effect (the increase in vocal effort when a speaker is in a noisy environment), speaker changes caused by sickness (cold), whispering, high g-load, etc. Distortions introduced by the environment and system are: noise, band-pass limiting, echoes, reverberation, non linear transfer, etc. Other causes of deterioration of the speech signal may be the use of an oxygen mask in an aircraft, co-channel interference, jamming, etc.

Humans can recognize speech under very poor signal-to-noise conditions (down to -15 dB, depending on the noise spectrum). In these extreme conditions only digits and the alphabet (alpha, bravo, etc.) can be understood correctly. A speech recognizer cannot handle such adverse conditions. In general the recognition performance decreases with signal-to-noise ratio. Optimal system development therefore requires an advanced signal treatment before recognition (i.e. noise cancelling microphones, speech enhancement units) or compensation mechanisms within the recognizer.

For automatic speech recognition we may reduce the requirements for the recognition procedure in order to gain performance under these limited conditions. For example speaker independent large vocabulary recognition requires a speech signal of high quality while speaker dependent small vocabulary recognition for isolated words is more robust with respect to deteriorated signals.

As an example, the results of an isolated-word recognition experiment are given in Table 3.1. The test vocabulary included 25 command and control application words used in an advanced aircraft cockpit. Several conditions were investigated with and without the use of an oxygen mask and with a variety of added noise levels. For the condition without the oxygen mask, the speech was recorded in a noisy environment (noise level of 80 dBA) and with noise added afterwards, i.e. the speaker was speaking in a silent environment. This procedure was repeated for the condition where the speaker was using an oxygen mask. A specific noise cancelling microphone was located inside the mask (different from the standard mask mike). As the sound attenuation of the mask is significant, the noise level was increased to 100 dBA and 110 dBA which represents the noise level of a moderate fighter cockpit. The recognition scores (percent correct) are given in Table 3.1.

Table 3.1. Mean recognition scores (mean % correct) and standard errors (se %) based on one speaker, ten test runs, seven noise conditions and with and without the oxygen mask. The noise was introduced in two ways, directly with the speaker placed in a noise environment (direct) and by addition after the recordings (additive).

Scores in % correct		Without oxygen mask			With oxygen mask			
		no noise	direct 80 dBA	add. 80 dBA	no noise	direct 100 dBA	add. 100 dBA	add. 110 dBA
	mean	100.0	98.0	84.0	93.6	83.6	86.6	61.2
	se	0.00	2.0	5.9	0.7	3.1	2.7	3.6

The results show that the addition of noise has a significant effect on the performance of the recognizer. Also the effect of the oxygen mask is significant. A similar performance was obtained for the additive noise condition 80 dBA without oxygen mask and 100 dBA with oxygen mask. One could say that the use of the oxygen mask improves performance in the noise conditions due

to the attenuation of the environmental noise. Optimal system design requires interaction between the applied front-end interfaces (e.g. microphone, signal processing) and the recognizer.

3.1.5 Speaker Recognition

3.1.5.1 Speaker Verification

Speaker verification is a method of confirming that a speaker is the same person he or she claims to be. It can be used, often in conjunction with other means such as passwords or security cards, as a means of access control to secure systems. It provides an additional confirmation of the identity of the user which cannot be stolen. The heart of the speaker verification system is an algorithm running on a computer or special purpose digital signal processor, which compares an utterance from the speaker with a model built from training utterances gathered from the authorized user during an enrolment phase. If the speech matches the model within some required tolerance threshold, the speaker is accepted as having the claimed identity. In order to protect against an intruder attempting to fool the system by making a recording of the voice of the authorized user, the verification system will usually prompt the speaker to say particular phrases, such as sequences of numbers which are selected to be different each time the user tries to gain entry. The speech verification system is combined with a recognition system to assure that the proper phrase was spoken. If it was not, the speaker will be rejected.

Depending on the application, different requirements can be imposed on the system depending on the perceived costs of having the system make a mistake. A system can make two types of errors. It can reject a legitimate user (false rejection) or it can admit an impostor (false acceptance). For any system, by adjusting the tolerance threshold, it is possible to make one of these parameters better at the expense of making the other worse. There are several parameters which determine the difficulty of speaker verification for a particular problem:

- Quality of Speech. In the ideal application, the speech can be gathered in a quiet environment with a high-quality microphone and transmitted to the verification system over a wide-band, high quality connection. This would be typical of a computer access control in which the user has direct access to the computer. In other cases, it may be necessary to access the system remotely by means of a telephone or other communications link. The collection and transmission of the speech may be significantly degraded making the verification problem much harder.

- Length of Utterance. Collecting a longer utterance will improve the performance up to a point, but will increase the time required for the test. There will be practical limits to what the user or the application will tolerate.

- Training Speech. Collecting more samples of speech from the user in multiple sessions over a longer period of time will improve performance by providing better models of the user. Over longer intervals it will also be necessary to update the training models as the speakers characteristics change.

There are a number of factors which should be considered in comparing systems based on different algorithms. These factors include computational requirements of the algorithm and the sensitivity of the algorithm performance to differences between training and test conditions. The changes may result from changes in the speaker, the environment, or the channel. Algorithms for speaker verification generally match the characteristics of the speech of the user with the speech collected during training. Often some normalization is done to compensate for changing conditions on the

channel or in the acoustic environment of the speaker. A survey of techniques has been prepared by Gish and Schmidt (1994).

Speaker verification is still an active area of research. Much progress has been made in the last few years. The US Department of Defence has developed a standard test corpus called YOHO which is designed for evaluating speaker verification systems. For these tests the performance standards were set as shown in Table 3.2.

Table 3.2. Speaker Verification Performance Specification.

		Requirement	Goal
Probability	(False Rejection)	0.01	0.001
	(False Acceptance)	0.001	0.0001

In first tests with this corpus, a few systems were able to meet the minimum requirements, but none has yet achieved the long-term performance goal (Campbell 1995).

3.1.5.2 Speaker Identification

Automatic speaker identification is the capability to identify a speaker from a group of speakers. It asks the question "Who does this speech sample belongs to?". This technology involves two steps: modelling the speech of the speaker population (training) and comparing the unknown speech to all of the speaker models (testing).

Speaker identification may be performed in a closed or open recognition mode. A closed set identification recognizer answers the question with the identity of an individual in the training set. That is, it forces the decision to be one of the speakers in the training set. An open set recognizer can identify speakers not in the training set (never seen or trained by the recognizer) as unknown. As with speaker verification, speaker identification performance is dependent on the speech quality, length of utterance and training conditions. Although active research is still required, recent advances in speaker identification make this technology of interest to military and law enforcement users in applications such as speaker tracking, surveillance, communications and other applications (Ricart et al. 1994).

3.1.6 Language Identification

A language identification system is used to identify the language of the speech coming over a channel of interest. Automatic language identification systems generally work by exploiting the fact that languages have different phoneme inventories, phoneme frequencies, and phoneme sequences. These features can be obtained, although imperfectly, using the same spectral analysis techniques developed for speech recognition and speaker recognition. The use of higher-level feature such as the prosodies and the use of expert knowledge about the target languages should also contribute to the language identification task, but to date, the best results have been obtained with systems which rely mainly on statistical analysis of spectral features.

There are practical issues which must be considered in putting together a system for a real application. Performance is of course a primary concern. This must be weighed against issues such as system complexity, the difficulty in training the system, and the ease with which new languages can be added. For example, the type and amount of data required for training could be very

important. Some systems can be trained given only examples of conversations in a given language. Others require detailed phonetically marked transcriptions in each language. The relative importance of these issues will differ depending on the constraints of the particular application. A recent survey article by Muthusamy (1994) describes many of the techniques.

The ultimate language identification system may be a set of parallel speech recognizers, one for each language of interest. Their outputs could be compared to determine which was the most plausible. Researchers are starting to experiment with such systems. However, the training problem is difficult. A complete recognizer must be trained for each language of interest.

Some of the factors which make the language identification problem easier or harder are the following:

- The quality of the speech and the channel over which it is received.
- The number of possible languages from which the system must choose.
- The length of the utterance on which the decision must be made.
- The amount of training data which is available for making the models for each language. Both total duration of the training speech and the number of different training speakers are important factors.
- The availability of transcripts of the training speech, text samples from the language, and phonetic dictionaries for the language to assist in the creation of the models for the language.

Language identification continues to be an area of research. A series of tests has been coordinated by NIST in the US over the last several years comparing the performance of language identification systems on a standard set of test data. The best results from the March 1995 test are summarized in Table 3.3 below. The test corpus contained 9 languages and was collected over long-distance telephone lines in the US by the Oregon Graduate Institute (OGI). The subject was recorded while talking to an automatic operator, which prompted the subject to speak about different topics. The results of two types of tests are shown in the table. In one case the system had to select one of the nine possible languages. In a second scenario, the system was asked to distinguish between pairs of languages consisting of English and one other language. Two utterance lengths were tested.

Table 3.3. Language Identification System Test Results.

(Percent Correct) Test Condition	Utterance Length	
	10s	50s
1 of 9 Forced Choice	77%	88%
Pairs Choice	96%	98%

Current areas of research include extensions to conversational speech, use over more difficult channels, and reducing the training requirements.

3.2 Language Processing

3.2.1 Topic Spotting

Topic spotting is a term used to describe a number of techniques used to monitor a stream of messages for those of particular interest. The messages may be speech or text messages. Typical applications include surveillance and screening of messages before referring to human operators. Closely related methods may be used for automatic literature searches, message prioritization and some simple 'natural language' database interfaces.

Topic spotting of speech messages may be based on whole words or phrases (word spotting), phoneme sequences or even acoustic feature vectors. Central to topic spotting is the concept of "usefulness". To be useful a feature (e.g. a word) must occur sufficiently often for reliable statistics to be gathered, and there must be significant differences in the distributions between the wanted and unwanted messages.

Simple topic spotting applications may be constructed using manually chosen key words, however more advanced applications generally process training data to automatically determine suitable features. Care must be taken to ensure the training data is representative of the target domain. Although systems for processing speech messages may be trained on transcriptions of the training data, better results are often obtained by processing the output of a suitable recognition system. The latter approach includes effects of detection errors and recognition difficulty. For speech based topic spotting, various recognition technologies may be used, ranging from large vocabulary recognition systems, phoneme transcribers or the use of special "general speech models" or "babble models" in conjunction with whole word recognition systems. While word or phrase based systems generally ignore details of the context of the features, some form of clustering and fuzzy matching techniques is often desirable in phoneme and acoustic feature based systems. The application, need for reconfigurability and available computer power may significantly constrain the techniques which may be employed. Often the systems need not operate in real time.

The choice of the unit used for detection will partially determine what distinguishes the types of message. When processing speech data the decision to base the system on words, phonemes or acoustic feature vectors will affect what the system is sensitive to. For example, a system based on phonemes may be sensitive to regional accents as well as certain words, while a word based system is more likely to be sensitive only to message content. Which is more useful will depend on the exact details of the application. Generally, to build up reliable statistics, a number of different features may be searched for an overall score for the message and is then based on the combined score. For continuous data, e.g. detection of weather forecasts in radio broadcasts, statistics are generally based on the frequency of occurrence of the features within some time interval. There may be a delay between the start of the section of interest and the output of the system.

Although topic spotting generally makes a binary decision - the message either is or is not of interest-, similar techniques may be used to classify messages into one of a number of categories. Performance of topic spotting systems are often described in terms of ROC (receiver operator curves) curves. These show detection probability as a function of false alarm rate.

3.2.2 Translation

A lot of excellent work has been performed within ARPA's TIPSTER Document detection and fact extraction, the MUC (Message Understanding) and the TREC (Text retrieval and Machine Translation) projects focusing on various tasks such as:

Evaluation of Machine Translation (MT) Systems

Metrics based evaluation techniques, particularly used in MT, have provided a somewhat more realistic but still too vague a basis for the user to apply ad-hoc to a generally unknown or black box MT system available on the market. It is, however, a solid base for assessing the performance of a development system.
Limiting the measurements to fluency (well-formed output), adequacy (accuracy and completeness) and informativeness (amount of information that is correctly conveyed) only still seems to underestimate the variance of competence-performance relationships based on a varying degree in quality and quantity of user input during the customizing phase.

3.2.3 Understanding

Speech system have progressed to the point at which they can begin to handle tasks which could broadly be described as language understanding. There is a tremendous range of problems which could fall under this definition. Systems are now starting to be applied to real, practical problems at the simplest end of this range.
Understanding problems can be divided into two broad categories. The first set of problems addresses human-machine interactions. In this case the person and the machine are working jointly to solve some particular problem. The interactive nature of the task gives the machine a chance to respond with a question when it does not understand the intentions of the user. The user can in turn, then rephrase the query or command. In the second type of problem, the machine has to extract some desired information from the speech without the opportunity for feedback or interaction.

The best examples of the human-machine interaction are found in some simple information retrieval systems. Such systems are beginning to move from laboratory prototypes into field demonstrations. Recently a spoken language systems program has used an Air Travel Information System as a model problem to drive the development of the technology. In this system, the user makes voice queries on an actual airline reservation database in order to make a plane reservation. Similar systems are being explored for use with actual reservations systems.

Another example is the prototype voice-driven telephone system which is used by the German Railways to provide schedule information for trains between cities in Germany. The system has a vocabulary of about 2000 words including 1200 station names. In spite of a 25% word error rate, the system is able to provide the correct response to more than 80% of the queries with less than 2 corrections.

An example of an experimental system which does not permit feedback is one used to monitor air traffic control messages (Rohlicek et al. 1992). This system tries to extract the flight identity from the voice message. Such a system has been considered as an aid to the air traffic controller. The system would identify the flight when a voice transmission is received and automatically highlight the information on the controller's display.

Further advances in understanding systems will build on progress in many fields, including speech recognition, natural language understanding, and man-machine interaction.

3.3 Interaction

3.3.1 Interactive Dialogue

A dialogue is usually considered to be an interchange between two cooperating partners during which some information is passed from one to the other. It may be better to treat the concept differently, recognizing that one of the partners has initiated the dialogue for a purpose that is not simply to conduct the dialogue. Accordingly, the two partners in a dialogue should be considered asymmetrically, one being the originator of the dialogue, the other being the recipient. The dialogue itself is successfully concluded when at least the originator believes that the recipient is in the state for which the dialogue was intended. The intended state may be that the recipient now has some information, or that the recipient has provided some information, or that the recipient is performing some task on behalf of the originator. In effect, a single one-way message has passed between originator and recipient, and has had a desired effect observable by the originator.

The back and forth flow associated with the common concept of "dialogue" consists of messages at a lower level of abstraction that enable the originator to ensure that the recipient receives the intended message (i.e. comes to a state that the originator intended). The dialogue "supports" the transmission of the one-way prime message. Each of these supporting messages, in turn, can be considered in the same way as the main message, as a one-way message that is to be supported by a two-way dialogue of messages at successively lower levels. What is commonly thought of as a "dialogue" can in this way be considered as a tree structure of messages in which each branch point represents a message in one direction or the other, and the leaves represent the physical realizations of the communication.

When the recipient is a computer, the same considerations apply as when the recipient is a human. The human originator wants to get the computer into some state, either to provide it with data, to get it to provide data, or to get it to perform some task. What the human wants the computer to do may not be something that can be done with one prepackaged command, and a dialogue is necessary. When the mode of the dialogue includes speech either as input to the computer or as output, or both, it is much less certain that the computer has correctly interpreted anything the human says than would be the case if the human entered the same words on a keyboard. Therefore there must be a possibility for a dialogue that supports the transmission of the individual utterances. Interpretation errors must be perceptible to the human as well as correctable when they occur.

Dialogue using speech I/O is not different in principle from dialogue with a computer using keyboard and display, but is valuable for different purposes and under different circumstances.

3.3.2 Multi-modal Communication

Speech is a natural mode of communication for humans, but, having evolved over thousands of years, language has become very complex. Machines are evolving too, but in the beginning they could utilise only very simple interfaces, such as switches. As the capabilities of computers have increased, more complex modes of communication with them have become possible, including the keyboard, the light pen and the mouse. Recently, computers have evolved to the point where the use of speech and natural language interfaces are also possible. The situation now is that several methods, or media, are available for human-computer interaction: this is described as a multi-media interface.

At present, the various media in a multi-media interface operate in parallel or as alternatives. Each mode of communication has its own sphere of operation. For example, a mouse is ideal for

moving the cursor to a particular place on the screen, but far from ideal for adding text at that point. The keyboard is very capable for adding text, but far from optimal for placing the cursor, especially if it has to be moved a long way. Considering a much more complicated example, a pilot could enter way point data into his navigation computer by voice, but would find switches more natural, or quicker, or more reliable, for other functions like lowering the undercarriage. Each interface mode is best suited to a particular kind of operation; matching the medium to the message is the essence of the art of interface design.

The next stage in this evolution is to make the various modalities interact and cooperate in carrying out tasks; this is the multi-modal interface. Now, each individual command to the system may utilise several interface modalities. This capability already exists in a limited form in the mouse, which combines pointing and a switch operation in a manner which seems natural to the user, or at least, is very easily to learn. The addition of speech to multi-modal interfaces will greatly increase their power. The combination of voice input and a pointing device would allow commands to use words like "this", "that" or "there", perhaps allowing a simplified display and removing the need for the operator to remember details of unit designations, etc. The associated displays and auditory outputs would need to be an integral part of such an interface, each tailored to provide information to the operator in the most easily assimilated form.

The problem that the multi-modal interface addresses is the mismatch between the increasing complexity of weapons systems and the more or less static capabilities of military personnel. The later can, of course, be improved by training, but this may be an expensive option, especially for those forces comprised partly of conscripts and thus having a rapid turnover of personnel. The alternative is to design the interface between the man and the machine in such a way that the machine comes to seem simple, or, at least, considerably less complex than it is. The combination of language (usually spoken but not always) and gesture is natural to humans and therefore minimises workload. A command such as "Put that over there", with the accompanying gestures, would be issued and understood by humans without significant effort.

The integration and fusion of information made possible by the multi-modal interface will allow the man-machine dialogue to be carried out at a higher level of abstraction than before. This is appropriate to the use of speech and language, as the words used may cover a wide range of levels of abstraction.

3.3.3 3-Dimensional Sound Display

Modern signal processing techniques allow headphone audio to be processed in such a way that it seems to originate from virtual sound sources located in the three-dimensional space around the listener. By using head tracking devices, it is even possible to create a stable virtual acoustic environment that takes (head) movements of the listener into account. One application of such a 3D-auditory display is enhancement of situational awareness by using virtual sound sources to indicate positions of relevant objects (e.g. targets or threats). Bronkhorst et al. (1996) have shown in a flight simulation experiment that 3D-sound can indeed be used as an effective acoustic radar display. They found that the time required to locate and track a target was similar for acoustic and visual displays and that the combination of both displays yielded even shorter search times.

A second, perhaps less obvious, application of a 3D-auditory display is its use as a means for improving communication and signal detection. This application is based on the ability of the human hearing system to tune in on sounds coming from one direction while suppressing sounds coming from other directions, a phenomenon which was coined as the "cocktail party effect". The results of a number of identification experiments clearly demonstrate this effect. One target speaker and up to four interfering speakers were presented simultaneously in three headphone conditions:

monaural, binaural without signal processing and binaural with 3D-sound. In the second condition, target and interfering speakers were divided over both ears; in the latter condition, the speakers were presented through virtual sources placed on a 180° arc in front of the listener. The results, obtained from six listeners, are plotted in Fig. 3.3. There appears to be a clear superiority of 3D-sound presentation over the other presentation modes.

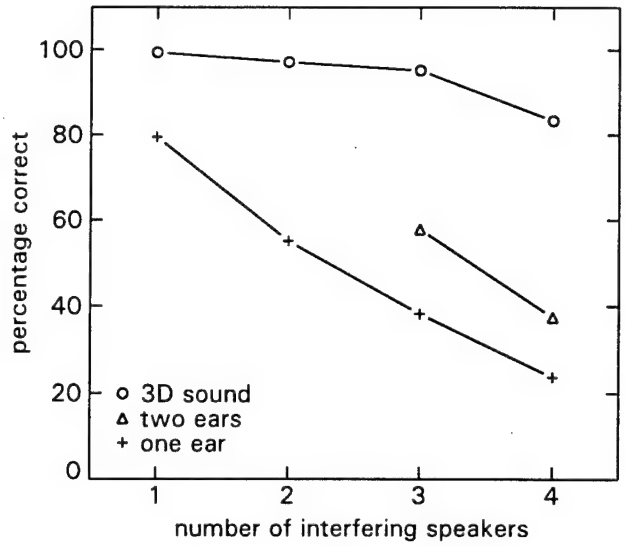


Fig. 3.3. Results of a number identification experiment, in which a 3D auditory display was compared with normal sound presentation.

A 3D-auditory display, therefore, offers significant advantages to persons who have to deal with multiple simultaneous signals. The "cocktail party effect" will help them to suppress unwanted sounds and the signals can also be presented from "meaningful" directions, to facilitate identification. Given the abundance of (military) functions where such situations occur (pilots, flight controllers, sonar operators, etc.), it seems probable that 3D auditory displays will become widely applied within the next few years.

Chapter 4. Assessment and Evaluation

4.1 Specification and Assessment of Speech and Language Processing Systems

4.1.1 Introduction

Very few speech and language processing applications involve "stand-alone" speech and language technology. Speech, handwriting and text provide essential components of the more general human computer interface alongside other input/output modalities such as pointing, imaging and graphics. This means that the actions and behaviours of the speech and language-specific components of a complex multi-modal "human-computer interface" (HCI) inevitably have to be orchestrated with respect to the other modalities and to the application itself, and this is usually achieved by some form of interactive dialogue process (simultaneously taking into account the wide range of human factors involved).

The complexity of the human-computer interface, and the subtle role of speech and language processing within it, has been (and continues to be) a prime source of difficulty in deploying speech and language systems in military applications. Not only are field conditions very different to laboratory conditions, but there has been a serious lack of agreed protocols for specifying such systems and for measuring their overall effectiveness.

This means that applications developers and system designers are unable to select appropriate "off-the-shelf" HCI components (such as automatic speech recognisers, for example) not simply due to a lack of standardised evaluation criteria for such system components, but also from a lack of a clear understanding of the implications of the performance of each system component on overall system effectiveness.

4.1.2 The right Application using the right Technology

One possible model for understanding the relationship between speech and language applications and the corresponding technology is illustrated in Fig. 4.1. The key notion is that, not only is it necessary to match the "capabilities" of the technology with the "requirements" of the applications (and this can be done at either the technical or operational levels), but it is also important to emphasise that the purpose of introducing speech and language technology into an application is to achieve the appropriate operational benefits.
The process illustrated in Fig. 4.1 shows how, in order to specify and assess the relevant technology, the operational benefits being sought in current and future speech and language applications (such as manpower savings or increased mission effectiveness) have to be expressed either in terms of operational requirements (such as increased data entry rates or reduced head-down time) or in terms of technical requirements (such as >200 words-per-minute data entry rate or $>95\%$ word recognition accuracy). Likewise the technical features of current and future speech and language technology (such as hidden Markov modelling or neural networks) have to be expressed in terms of the corresponding technical or operational capabilities.

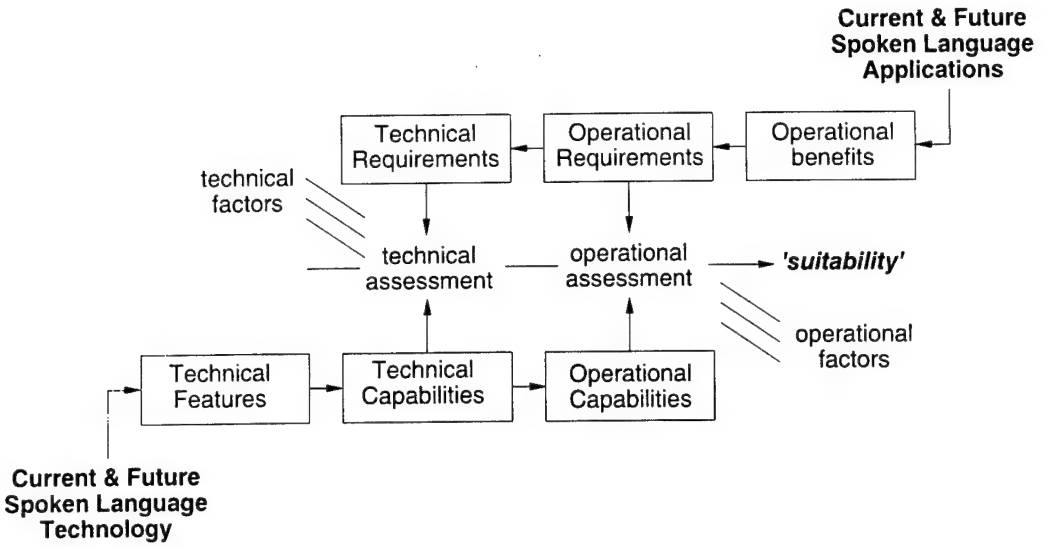


Fig. 4.1. A model of the relationship between the applications of speech and language systems and the underpinning technology.

4.1.3 System Specification

Of course many different factors influence the suitability of particular speech and language systems for specific applications (such as the level of acoustic noise in the environment or the degree of training which the user has received), therefore the requirements and capabilities are expressed as multi-dimensional "profiles"; the "capability profiles" indicates the performance that is available (what can be done) and the "requirement profiles" indicates the performance that is required (what is needed).
As a consequence, the specification of a speech and language system inevitably involves a process whereby a system designer would have to provide qualified judgements about the required values and acceptable ranges (together with the weighting of each dimension in terms of its relative importance) against a comprehensive checklist of performance-related factors.

For example, for an automatic speech recognition system the list of technical requirements would have to specify the characteristics of all of the following influencing factors:

- environment
- transducer
- channel
- task
- talkers(s)
- talking style
- enrolment
- implementation
- controls

and also the following performance requirements:

- recognition error rate
- speed
- response time
- enrolment time
- size
- packaging
- weight
- power
- cost

4.1.4 Evaluation and Assessment

The overall goodness of a speech and language system can be viewed as corresponding to a "performance envelope" in which performance in one dimension can be traded against

performance in another. However, in many circumstances the most appropriate technology may not be available for a given application, and this leads to the concept of a best fit between requirements and capabilities (where some requirements may not be satisfied). Shortfalls in one dimension would have to be traded against gains in others, and adjustments to the capabilities (or indeed the requirements) would have to be made in order to fulfil some bottom-line criterion (such as minimum cost, for example).

In the assessment of speech and language systems it is possible to distinguish three main methodologies:

- live "field" trials
- laboratory-based tests
- system modelling paradigms.

The first of these of course is likely to provide the most representative results but, from a scientific point of view, there are likely to be a number of uncontrolled conditions and this limits the degree of generalisation that can be made from application to application. Field trials also tend to be rather costly operations to mount. Laboratory testing is per force more controlled and can be relatively inexpensive, but the main problem is that such tests may be unrepresentative of some (possibly unknown) key field conditions and give rise to the observed large difference between performance in the laboratory and performance in the field. The third possibility, which is itself still the subject of research, is to model the system (and its components) parametrically. In principle, this approach could provide for a controlled, representative and inexpensive methodology for assessment but, as yet, this area is not sufficiently well developed to be useful.

The term "assessment" also covers a range of different activities. For example, a suitable taxonomy of assessment activities should include: "calibration" (does the system perform as it should), "diagnosis" (how well does the system perform under parametrically controlled conditions), "characterisation" (how well does the system perform over a range of diagnostic conditions), "prediction" (how well will the system perform under different conditions) and "evaluation" (how well does the system perform overall). Of all these, the last evaluation has received a bulk of the attention in speech and language systems assessment.

It is also the case that assessment protocols are required which address a large number of different types of speech and language systems. For example, such systems range from laboratory prototypes to commercial off-the-shelf products, from on-line to off-line systems, from standalone to embedded systems, from sub-systems to whole systems and from speech and language systems to speech and language based HCI systems.

The majority of activity in the area of speech and language system assessment has concentrated on evaluating system components (such as measuring the word recognition accuracy for an automatic speech recogniser, for example) rather than overall (operational) effectiveness measures of complete HCI systems. Since the publication of the US National Bureau of Standards guidelines in 1985, there have been considerable developments at the international level. In Europe, the ESPRIT Speech Assessment Methods project established a standard test harness for both recognisers and synthesizers and in the US a very effective assessment paradigm has been funded by ARPA which included an efficient production line of "hub and spoke"-style experiments involving the coordinated design, production and verification of data, distribution through the Linguistic Data Consortium, and with the National Institute for Science and Technology responsible for the design and administration of tests and the collation and analysis of the results.

These activities emphasis the importance of "bench marking", either through the implementation of standard tests, or by reference to human performance or to reference algorithms.

4.1.5 Standards and Resources

The requirement for agreed standards and guidelines pervades all of the links in the speech and language system R&D chain starting from the research community (for algorithm development and benchmarking), to product developers (for performance optimisation), system integrators (for component selection), manufacturers (for quality assurance), sales staff (for marketing), customers (for product selection) and users (for service selection).

In addition, many of the speech and language technologies rely heavily on the availability of substantial quantities of speech and language corpora: first, as a source of material from which to derive the parameters of the constituent models, and second, in order to assess performance under controlled (repeatable) test conditions.

The most significant activity on speech and language standards and resources has been the ESPRIT Speech Assessment Methods (SAM) project which ran from 1987 to 1993. The SAM project arose out of the need to develop a common methodology and standards for the assessment of speech technology systems which could be applied within the framework of the different European languages.

4.1.5.1 Corpora

Three types of speech and language corpora are typically of interest: "analytic-diagnostic" material which is of primary importance to progress in basic science and which is specifically designed to illuminate specific phonetic and linguistic behaviour, "general purpose" material which includes vocabularies which are either common or which are typical of a wide range of applications (for example, alpha-numeric words or standard control terms), and "task-specific" material which reflects different levels of formalised spoken monologue/dialogue within constrained discourse domains.

Clearly general purpose corpora are easy to collect and are useful in a general sense but, of course, they have only limited practical value. On the other hand, although task-specific corpora can be time-consuming to collect and are only relevant to a specific domain, they are obviously directly useful for the purposes of practical applications. Diagnostic corpora are time consuming to design, but they are extremely useful for research purposes.

The availability of standard corpora is of great importance for the speech community and a number of national and international bodies have been responsible for co-ordination, distribution and production of appropriate databases.

For military applications, NATO Research Study Group on Speech Processing (AC342/Panel 3/ RSG.10) has, since the late 1970s, provided an effective mechanism for exchanging information on spoken language standards and resources between Canada, France, Germany, the Netherlands, the UK and the USA. RSG.10 was responsible for the first publicly available multi-lingual speech corpus, and has subsequently released on CD-ROM a database of noises from a range of selected military and civil environments (NOISE-ROM) and related experimental test data (NOISEX). In addition RSG.10 maintains a database of speech corpora specifically related to the military needs. At time of writing the database contains about 45 entries, some of which include non-speech sounds such as background noise in military vehicles.

4.1.5.2 Other resources.

In the US, the Linguistic Data Consortium (LDC) was founded in 1992 to provide a new mechanism for large-scale development and widespread sharing of resources for research in linguistic technologies. Funded with an initial grant of $5 million from ARPA and membership fees from over 65 companies, universities, and government agencies, the consortia distributes previously-created databases, as well as funding and coordinating the funding of new ones. The LDC is closely tied to the evolving needs of the community it supports and has helped researchers in several countries to publish and distribute databases that would not otherwise have been released.

In Europe, the European Language Resources Association (ELRA) was established as a non-profit organization in 1995, with the goal of creating an organization to promote the creation, verification, and distribution of language resources in Europe. Eventually, ELRA will serve as the European repository for EU-funded language resources and interact with similar bodies in other parts of the world (such as the LDC).

Another valuable resource is a series of handbooks of speech and language standards and resources produced as a result of an initiative by the Commission of the European Union under the auspices of the DG XIII Linguistic Research and Engineering (LRE) Programme. The initiative became known as "EAGLES" the Expert Advisory Group on Language Engineering Standards. The initiative covers a wide range of topics including methodologies for the creation and interchange of electronic language resources such as text and speech corpora, computational lexicons and grammars formalisms, and the evaluation and quality assessment of language processing systems and components.
The spoken language handbook, completed in 1996 covers: the design and specification of speech and language systems; speech and language resources (the design, collection, characterisation and annotation of corpora, lexica and language models); assessment methods (for recognition, synthesis, verification, and interactive systems); as well as containing a substantial body of reference material.

4.2 Specification and Assessment of Speech Communication Systems

4.2.1 Introduction

Evaluation of speech related aspects of communication systems can be separated in intelligibility, response in adverse conditions (jamming, channel noise, background noise, etc.) and the response time (delay). Generally the intelligibility or speech quality is measured under various representative conditions. Criteria for the performance under optimal conditions (back-to-back connections) and representative usage conditions are proposed. Measures of performance and intelligibility can be determined both by subjective and objective methods. There are three groups of measuring methods:

- subjective intelligibility measures based on phonemes, words or sentences,
- subjective quality measures related to a global impression, and
- objective measures based on physical aspects of the speech signal or the speech transmission system.

Subjective intelligibility measures are in general very representative as speakers and listeners are used. However, to obtain reproducible results, much effort is required to perform the measurements and results are dependent on the speech material used for the test. Also no diagnostic information is obtained. One simple method is based on quality rating where listeners

score their impression of the speech quality. This is a global method and requires many listeners. Finally objective methods were developed in which the transmission quality is derived from physical parameters. These methods are easier to apply and offer additional to the prediction of intelligibility also useful diagnostic information. Unfortunately these methods cannot be used for voice coders like LPC-based systems.

4.2.2 Intelligibility Measures and Quality Rating

A number of subjective intelligibility tests have been developed for the evaluation of speech communication channels. In general, the choice of the test is related to the purpose of the study: are systems to be compared or rank-ordered, are systems to be evaluated for a *specific application* or must the *development* of a system be supported. For both types of application a different test may be appropriate. An overview focused on the assessment of speech processing systems is given by Steeneken (1992).

Subjective intelligibility tests can be largely categorised by the speech items tested and by the response procedure used. The smallest items tested are at the segmental level, i.e. phonemes. Other test items are CV, VC, and CVC combinations (C=consonant, V=vowel), nonsense words, meaningful words, and sentences.

Besides intelligibility scores, speech quality can also be determined by questionnaires or scaling methods, using one or more subjective scales such as: overall impression, naturalness, noisiness, clarity, etc. Speech quality assessment is normally used for communications with a high intelligibility, since most tests based on intelligibility scores are inappropriate because of ceiling effects.

Tests at phoneme and word level

A commonly used test for determining phoneme scores is the rhyme test. A rhyme test is a forced-choice test in which a listener, after each word that is presented, has to select his response from a small group of visually presented alternatives. In general, the alternatives only differ with respect to the phoneme at one particular position in the test word. For example, for the Dutch language and for a test with a plosive in the initial consonant position, the possible alternatives might be: Bam, Dam, Gam, Pam, Tam, Kam. A rhyme test is easy to apply and does not require much training of the listeners. Frequently used rhyme tests are the Modified Rhyme Test (MRT, testing consonants and vowels) and the Diagnostic Rhyme Test (DRT, testing specific initial consonant pairs only).

A more general approach is obtained with a test with an *open* response, such as with monosyllabic word tests. Open response tests make use of short nonsense or meaningful words most often of the CVC type.
The test results can be presented as phoneme scores and word scores but also as confusions between the initial consonants, vowels, and final consonants.
The confusion matrices obtained with open response tests provide useful (diagnostic) information for improving the performance of a system.

Tests at sentence level

Sentence intelligibility is sometimes measured by asking the subjects to *estimate* the percentage of words correctly heard on a 0−100% scale. This scoring method tends to give a wide spread among listeners. Sentence intelligibility saturates to 100% at poor signal-to-noise ratios, the effective range is small (see Fig. 4.2).

Quality Rating

Quality rating is a more general method, used to evaluate the user's acceptance of a transmission channel or speech output system. For quality ratings, normal test sentences or a free conversation are used to obtain the listener's impression. The listener is asked to rate his impression on a subjective scale such as the five-point scale: bad, poor, fair, good, and excellent. Different types of scales are used, including: intelligibility, quality, acceptability, naturalness etc. Quality rating or the so-called Mean Opinion Score (MOS) gives a wide variation among listener scores. The MOS does not give an absolute measure since the scales used by the listeners are not calibrated. Therefore the MOS can be used only for rank-ordering conditions. For a more absolute evaluation, the use of reference conditions is required as a control.

Objective Intelligibility Measures

The first description of the use of a computational method for the prediction of the intelligibility of speech and its realization in an objective measuring device, was developed in 1959 by Licklider. Presently a measure based on the Speech Transmission Index (STI, Steeneken and Houtgast 1980) is standardized by IEC 268-16.

The method assumes that the intelligibility of a transmitted speech signal is related to the preservation of the original spectral differences between the speech sounds. These spectral differences may be reduced by band pass limiting, masking by noise, nonlinear distortion components, and distortion in the time domain (echoes, reverberation). The reduction of these spectral differences can be quantified by the effective signal-to-noise ratio, obtained for a number of relevant frequency bands. As the STI is focused on the reproducibility of the spectral and temporal *envelope* and does not take into account the reproducibility of the carrier, the method cannot be applied to vocoders.

For the application of the STI method a specific test signal is applied at the input side of the system under test. An analysis is made of the output in order to obtain the effective signal-to-noise ratios for all frequency bands considered (seven octave-bands ranging from 125 Hz to 8 kHz). The test signal and the analysis is designed in such a way that nonlinear distortion and distortion in the time domain affect the information content of the test signal in a manner similar to the degradation of speech. A weighted contribution of the measured information transfer in the seven octave bands results in a single index, the STI. The measuring method and the algorithm have been optimized for an optimal correlation of the STI with the subjective intelligibility.

Criteria and Relation between Various Measures

Fig. 4.2 shows, for five subjective intelligibility measures, a quality rating and the relationship with the objective STI. Also shown are comparable signal-to-noise ratios. This illustrates the effective range of each test. The given relation between intelligibility scores and the signal-to-noise ratio is valid only for noise with a frequency spectrum similar to the long-term speech spectrum. In this instance a voice-babble is used. A signal-to-noise ratio of 0 dB then means that speech and noise have an equal spectral density.

As can be seen from the figure, the CVC-nonsense words discriminate over a wide range, while meaningful test words have a slightly smaller range. The digits and the alphabet give a saturation at a signal-to-noise ratio of -5 dB. This is due to: (a) the limited number of test words and (b) the fact that recognition of these words is controlled mainly by the vowels rather than by the consonants.

In general for military communications a back-to-back performance qualified as good is required. This corresponds with a minimum CVC-word score of 70%, a DRT of 96% or an STI of 0.6. The criterium for just acceptable in the worst condition and qualified as poor (e.g. related to a condition that less than 100% intelligibility of digits and redundant sentences is obtained) corresponds to a CVC-word score of 40%, a DRT of 76% or a STI of 0.35.

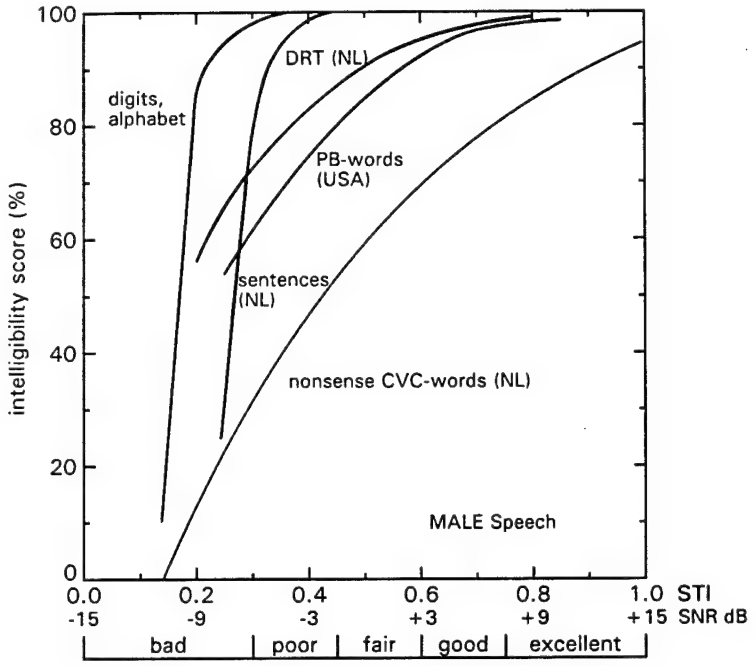


Fig. 4.2. Relationship between some intelligibility measures according to Steeneken (1992). Also shown is the relation with signal-to-noise ratio. (Noise with a spectrum shaped according the long-term speech spectrum.)

Chapter 5. Case Studies and (future) Applications

In this chapter, applications of speech and language technology that are already in service or likely to be in the near future are presented. The selection of contributions is not intended to be exhaustive, but shows the level of maturity of the main technology areas and covers a wide spread of applications. Automatic speech recognition frequently draws the main attention, but it should not be forgotten that coding, synthesis and other technologies also have their applications, and indeed are more advanced in terms of actual use. Secure speech transmission using LPC-10 at 2400 Bps has been in use for many years and saw active service in the Gulf War. Similarly, voice warnings are in use in many aircraft cockpits, both military and civil.

Speech recognition has perhaps the widest range of possible applications. Its use is almost imperative in modern single-seat fighter and strike aircraft, where system complexity makes the workload very high. It finds broader use as part of a multi-modal interface in operations room applications, which might be aboard ship, on land or in the air. The object is to reduce the operator's workload or to enable him to handle more data. In training situations, such as for air-traffic controllers, the main motivation is the saving of the time of highly trained personnel who would otherwise have to play the parts of the aircraft being controlled.

Similar workload and training considerations make voice technology particularly attractive for space operations. To meet the increasingly demanding space mission objectives, astronauts must undergo extensive training and learn to interact with a myriad of different on-board systems. Constantly evolving, the spacecraft operational environment can only benefit from added interactivity.

In the communications area, reductions in the data rate required for speech transmission will aid covert operations, or allow secure, long range communication with mobile units over HF radio. Maintenance of the quality and intelligibility of the speech is vital, as are suitable means of measuring these factors. The applications described below are mostly demonstrations or experimental systems; in several cases, however, it appears that only the will and the money would be required to put the systems into service. The technology now has adequate performance for useful application in many areas.

5.1 Cockpit Fast Jet

The ever-increasing complexity of aircraft systems coupled with requirements to operate at very low level and in all weathers creates a high workload in military cockpits, especially in single seat aircraft. A pilot's top priority should be to fly the aircraft, which requires the use of his hands and eyes. The operation of other equipment, although necessary for the mission, may be a distraction from the primary task. It has long been recognised that automatic speech recognition could provide a means of alleviating the workload and increasing eyes-out and hands-on time. Research suggests that this could have a significant impact on safety and mission effectiveness.

The military cockpit is however a very difficult environment for a speech recogniser. There is a high level of background noise and many factors which cause variations in the pilot's voice. On the other hand, there is a requirement for a very high level of accuracy: the pilot must be confident that the aircraft systems will respond as he desires. This has created an impression in some quarters that the technology will never be good enough, but progress continues to be made. For example, special algorithms can be used for recognition in high levels of noise, achieving near 100% accuracy at 0 dB speech-to-noise ratio. Other developments are making recognisers more robust to variations in speech.

Successful implementation of automatic speech recognition in fast-jet cockpits will require careful attention to the human factors aspects. Not all tasks in the cockpit are suited to voice input. Once the appropriate tasks have been chosen, the vocabulary and syntax must be designed, and suitable feedback methods implemented. Above all, voice input must be regarded as an integral part of the cockpit design, and not as an optional extra. Recent results from trials in several countries indicate that the technology can deliver adequate performance for the more consistent speakers in less demanding flight conditions. There are reasonable grounds for expecting that the performance envelope can be expanded sufficiently to realise the benefits outlined above.

For example, a research project is being conducted on Automatic Speech Recognition applications in a state-of-the-art single seat fighter cockpit in the Netherlands. Applications have been developed for integration in an F-16 cockpit in a Mid-Life Update (MLU) configuration. Cockpit tasks which can be executed with ASR include enhanced data entry operation for the Communication, Navigation and Identification (CNI) systems, and display management and control functions. Furthermore, the ASR applications include tasks supporting the normal Hands-On-Throttle-And-Stick (HOTAS) concept and interactive "crew assistant" applications using voice feedback. The primary pilot feedback mechanism is a normal avionics system response. A Press-To-Talk (PTT) switch is used to address the ASR system. The syntax and vocabulary have been developed on a commercial continuous speech recognizer; the ASR application has over 250 vocabulary words, 300 nodes and more than 4,000 node-to-node connections.

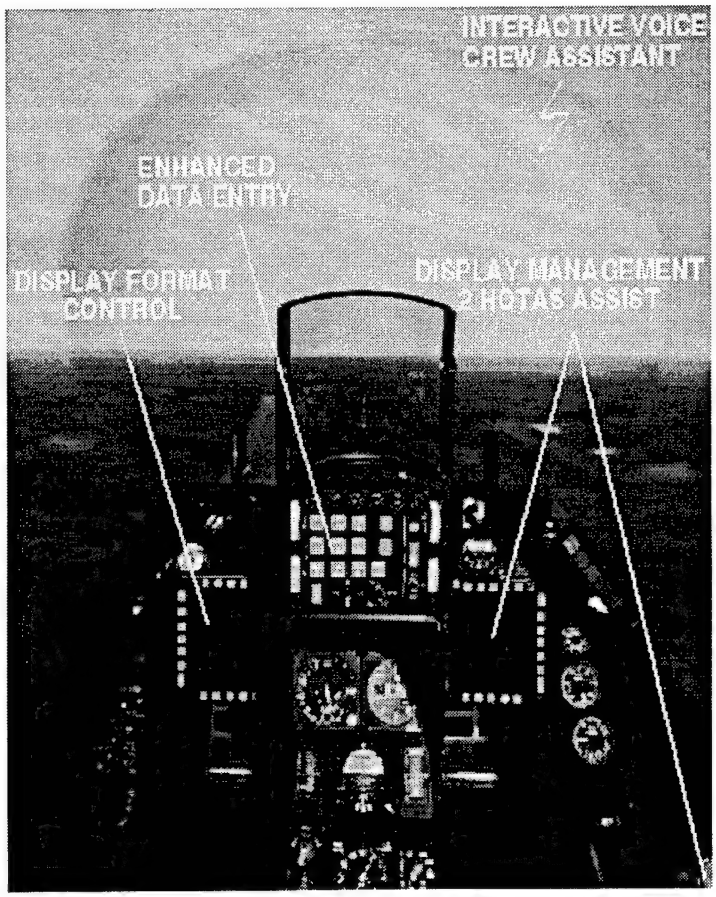


Fig. 5.1. Set-up of voice controlled F-16 simulator cockpit at the National Aerospace Laboratory in the Netherlands.

The integrated application will be evaluated on a MLU F-16 simulator (see Fig. 5.1) with a completely functional cockpit, a head-slaved outside visual dome projection system with a background and a high resolution area-of-interest inset providing the pilot with a "full" visual environment, a sound system, and a six-degree-of-freedom motion system. Simulator evaluations were conducted in operational scenarios to investigate integration aspects, such as task suitability for ASR, performance under practical conditions, pilot acceptance and pilot attention distribution aspects.

5.2 Helicopter

Military helicopters, as well as other military aircraft, become more and more complex because of the increased capability to perform a complex mission. Recent extensions have included more sophisticated sensors and diversified weaponry. A modern helicopter is a part of a complex system in which also command posts, armoured vehicles, other aircraft, and real-time intelligence centres work together.

The complexity of the systems embedded into the aircraft implies development of new concepts of the Man-Machine Interface. The tactical piloting task, under adverse conditions (stress, night, bad weather, vibration, and noise), gives a high workload for the crew. For this reason the interfacing to systems within the cockpit has to be simple. This also will prevent operating errors.

In this framework, at Sextant Avionique of France, a speech recognizer is applied in a tactical management system. The system includes an on-board station and a ground station. Both stations share a common mapping of the local area. This implies the management of a cartographic database and of additional objects which are related to the military units in the area.

By using a specific speech recognition system (Topvoice), the man-machine interfacing of the embedded system was drastically simplified. All operations dealing with visualisation and configuration (zooming, zone-to-be-displayed, map layers display), and the operations for tactical object management (add, move, delete, tagging) are made by direct voice input.

Numerous tests during in-flight conditions have been completed in 1995 by more than twenty different speakers. The performance of the voice input system was 95% correct responses to the spoken commands even under adverse environmental conditions (windows open and maximum engine power).

The application has proven that the concept of voice input for military helicopters can be extended to others systems which require discrete operating control:

- System interface such as inquiries to get status from sub-systems of the aircraft (engine status, fuel management),
- Sensor suite management,
- Non-decisive actions concerning the weaponry.

5.3 Sonar

In the beginning sonar systems were very simple as described for example by Leonardo da Vinci: "If you place the head of a long tube in the water and place the other extremity to your ear, you will hear ships at a great distance from you." Since that time, the science of underwater acoustics has been refined and is still progressing rapidly. Today more and more often the complexity level

and the amount of information available at a sonar suite output are exceeding the capacity limits of the classical man-machine interface. The main issues are:

- the greater detection range capabilities (sometimes more than the radar range) and consequently the increased number of detected contacts;
- the increased amount of information extracted for each contact (position, broad band and narrow band spectral description, transients, intercepted active sonar pulses);
- the greater algorithm complexity and consequently the increasing number of parameters;
- and finally the increasing number of graphical interactive tools aiming to help sonar operators in their various tasks. This includes contact tracking, data fusion, classification, contact motion analysis (in particular in the bearing only passive mode), situation and threat assessment (for a ship or for a zone), decision about manoeuvres, weapons and sonar suite use.

It is worth noting that although high level information is automatically provided by the sonar system, the interpretation by sonar analysts is still frequently required. For example analysis of low level spatiotemporal signals. Finally in the ASW domain (anti-submarine warfare) everything is slower than in the air defence domain but paradoxically that does not really simplify the operators task because they must take into account all the various information gathered during a rather long period (a few hours).

Currently on board submarines, surface ships or maritime patrol aircraft, the sonar operators are overloaded. Because of the high training level and skills required for analysts, it is costly to increase the number of operators. Two solutions may be feasible to improve the performance of the operators by improving the various sonar analysis algorithms and by improving the efficiency of the man-machine interaction. This last point constitutes an opportunity for using man-machine communication by voice, integrated to the use of the standard display, keyboard and trackball (or joystick).

At present, voice input is not used operationally in the ASW domain. However, experimental studies have been carried out to integrate the speech recognizer into the more general man-machine interaction system. Voice input can be integrated in the following tasks:

- Panoramic surveillance on board submarines or ASW surface ships. This includes management of the intermittent tracks and contacts provided by the sonar suite, association/fusion of tracks, and recording of information about each contact (such as acoustic data, behaviour, classification, crossing with other tracks). Speech recognition can also be useful for controlling the interactive display. For example requesting the intermediate results from raw signals, detected events, lofar, etc.

- Classification tasks performed on board ASW ships or aircraft or in shore-based intelligence centres. Tools for the analysis of the contact signature and the matching to known submarine or surface ship signatures are to be controlled by voice. Voice input can also be useful for controlling the display itself.

- Control of the various interactive graphical tools used for situation and threat assessment, and the decision aids concerning manoeuvres, weapons and sensors.

- ASW tactical training systems to be used in land-based training centres.

In all these sonar related tasks, voice input based on connected word recognition provides the following benefits:

- The user can look continuously at the object without looking at menus or the keyboard.

- The cursor remains available for pointing out objects on the screen rather than pointing out menu items.
- The screen surface allocated to menus is reduced.
- Consequently the interaction can be faster.

These advantages become more important as the Control Information and Command (CIC) room has generally a low light level and as the use of the cursor (controlled by the trackball) and of the keyboard can be difficult in this situation especially when the ship is rolling and pitching.
It is worth noting that the CIC layout will probably be slightly modified by the introduction of man-machine communication by voice: in order to avoid disturbing the sonar analysts by the spoken orders given by other operators, headphones must be provided to every analyst and the major part of the human-human communication must be transmitted by headphones. Another possible solution to this problem is to move the consoles away from each other.

5.4 Noise Reduction

GRYPHON is a real-time noise reduction system which runs on UNIX workstations. GRYPHON integrates different noise reduction algorithms and audio manipulation functions making it a useful multi-purpose noise reduction tool. GRYPHON can be controlled interactively by operators or utilized in a fully automatic mode.
The system is currently used by the Canadian Department of National Defence to:
- reduce tonal and broadband noises in digital signals,
- reduce listeners' fatigue due to noisy voice communication channels,
- increase the intelligibility of speech signals degraded by coloured broadband noise, and
- improve the performance of automatic speech recognizers and speaker recognition systems.

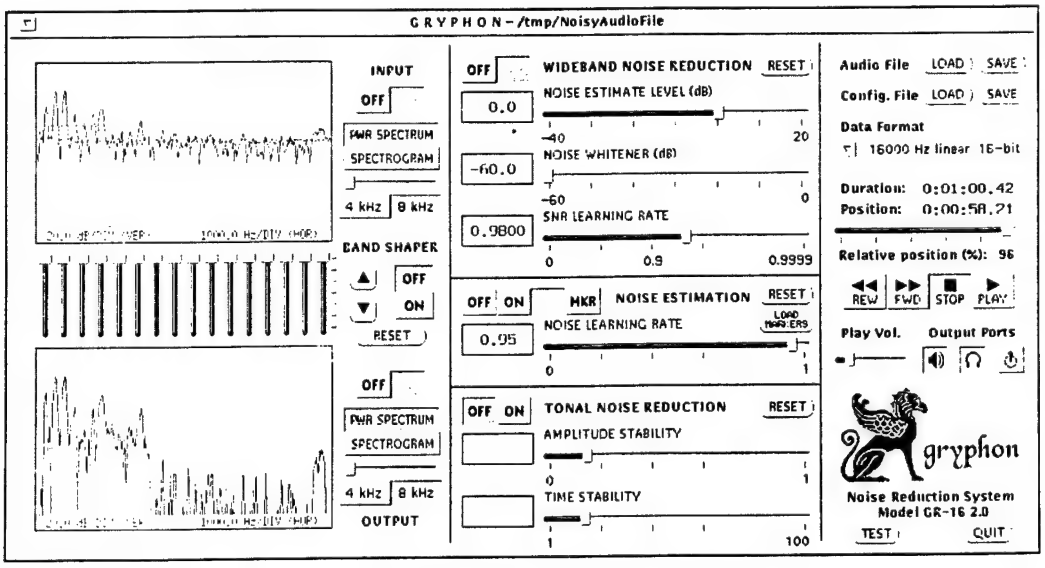


Fig. 5.2. The GRYPHON noise reduction system.

The system has these characteristics:
- wide-band noise reduction,
- semi-automatic training or utilization of labels to model noise interference,
- tonal noise reduction,

- user-friendly audio file playback control,
- linear-band shaping for signal equalization, (e.g. to improve the frequency response),
- real-time power spectrum or spectrogram displays,
- test signal generator,
- ability to save enhanced signal files to disk for further processing and playback,
- ability to save system status and noise models to disk for later use.

5.5 Training of air traffic controllers

Air traffic controllers are required to give clear verbal instructions to pilots and are trained to use a very limited English grammar and phraseology. A number of groups have invested significant research effort into finding possible uses for speech recognition technology, both for operational and training use. Many national civil aviation authorities have research programmes and several have issued formal specifications for training systems using speech recognition technology. A number of commercial systems are in the advanced stages of development.

At least three potential applications have been identified so far. The simplest application is to monitor the controllers speech for aircraft's call signs. When identified, flight strip information for the aircraft is activated to allow the controller to inspect or modify it. Since at any time the number of call signs being used by the controller is quite small, it is possible to build a reasonably reliable system without requiring any special co-operation from the controller. Such systems can provide considerable operational advantages and help reduce the work load on the controller.

A more ambitious use of recognition technology is to help automate the work of pseudo pilots during training. At present, during training, pseudo pilots listens to the trainee controllers commands and enter the commands into a simulator and give feedback to the controller. During some stages of training more than one pseudo pilot is required per trainee. To provide a fully automatic system requires the recognition of more than a hundred phrases with a good recognition accuracy. Systems with smaller phraseology may be used during early stages of training, or for self study exercises. Since a full training course can require many hundred hours of practice, use of speaker dependent systems is acceptable. In some trials speaker independent systems have provided better results, particularly since the controllers voice can vary significantly during training, or when under stress. While the most obvious advantage of the use of speech recognition technology is to reduce the need for pseudo pilots and so offer more opportunity for practice, incidental advantages include training the controllers to adhere more closely to the textbook phraseology and to adopt a clear and consistent speech style. It is also possible to design systems which automatically log mistakes for later analysis and debriefing.

A third use of recognition technology is when training controllers whose first language is not English. This is a particularly acute problem in Europe. Although such controllers are generally given general purpose English (Weinstein 1994) language training before their operational training begins, the training often does not give sufficient practice for the highly specific air traffic control vocabulary. A number of multi-media systems incorporating speech recognition technology may be used for phraseology training.

5.6 ARPA Spoken Language Systems Demonstrations and Applications

Over the past decade, the Advanced Research Projects Agency (ARPA) in the United States has played a crucial role in the research and development of spoken language technology. Tremendous advances have been made in both speech recognition and speech understanding, which have

created unprecedented opportunities for major improvements in the effectiveness of human-computer interactions in military, government, and commercial systems.

In April 1993, ARPA sponsored a special workshop on Spoken Language Technology and Applications (Weinstein 1994), which featured a broad range of live demonstrations of spoken language systems applicable to military systems.
The demonstrations and related applications included:

DEMONSTRATION	APPLICATION
Voice-Controlled Flight Simulator	Voice control of Aircraft Instruments
Recognition of Aircraft Identification	Voice Access to Air Traffic Control Data
Language Education with Speech	Training of Special Forces
Large-Vocabulary Recognition/Dictation	Document and Report Creation
On-Line Air Travel Planning	Travel Planning
Gisting of Voice Traffic	Identification of Flights & Events for Air Traffic Control
Direction Assistance via Voice	Route-Finding & Situation Awareness

The demonstrations were followed by a user panel which identified a wide range of military applications, including command-and-control on the move, tactical display control and data entry, multi-modal input/output in the cockpit, and training of air traffic controllers for both civilian and military operations.

In August 1995, a new set of demonstrations was shown at an ARPA Software Technology and Intelligent Systems Symposium. The 1995 demonstrations were both more advanced, and more directly focused on military applications. These demonstrations are described here, as an indication of the kinds of spoken language systems applications for the military which are possible within the next few years.

Informedia: News-on-Demand

Informedia News-on-Demand is a demonstration system developed at Carnegie Mellon University, which applies two key speech technologies: (1) automatic monitoring of the voice track of TV and radio news; and (2) selective retrieval of news items based on queries spoken by the user. Speech recognition helps to create time-aligned transcripts of spoken words; even though the word recognition accuracy for this difficult, open-vocabulary task is not very high, enough words can be identified to produce useful selective retrieval of news items of interest. In addition, when closed-caption text is available, the speech recognition helps to align that text with the audio track of the news broadcast. During exploration of news events by the user, speech recognition is employed to allow direct interaction with the system by voice. Informedia demonstrates the capability to allow users to navigate efficiently through the complex information space of news stories, and is a model demonstration for systems that will allow convenient, selective voice-controlled information retrieval from multi-media sources.

SpeechWear

SpeechWear, developed by Carnegie Mellon University, is a wearable computer that users can talk to, allowing hands-free operation. It is being developed initially for amphibious vehicle maintenance by United States Marine Corps personnel. The Speech Wear interface consists of a speech-enabled World-Wide Web type of browser that allows the user to navigate manuals and forms. Using a small head-mounted display, the operator can obtain maintenance information in

useful form while keeping hands on the equipment. The SpeechWear paradigm is readily extensible to other domains.

GALAXY: Accessing the information highway using spoken language

The GALAXY system, developed at the MIT Laboratory for Computer Science, enables information access to a wide variety of sources in the world-wide Information Infrastructure, using natural spoken dialogue. The system is currently connected to many on-line databases, including commercial air travel information networks, national weather services, electronic yellow pages, and the World Wide Web, and it is easily extensible to other domains. Users can query the system in natural English (e.g., "what is the weather forecast for Miami tomorrow", "do you have any information on the Olympic Games", etc.) and receive verbal and visual responses. GALAXY integrates several spoken language technologies to achieve its goals. On the input side, speech recognition and natural language processing are combined to derive an understanding of the input, often in the context of the ongoing dialogue. On the output side, natural language processing and speech synthesis are combined to convert the information that the user seeks, as well as any necessary clarification dialogue, into natural sentences which are delivered as verbal responses.

CommandTalk: spoken language interface to the LeatherNet system

CommandTalk, developed by SRI International, is a spoken language interface to an advanced simulation and training system called LeatherNet, which is being developed for the United States Marine Corps. The setup and control of an advanced simulation system and its components is an important and complex operation, which requires a good deal of facility with graphical and keyboard interfaces. CommandTalk provides a spoken language capability to help make this facility easier to learn and use. CommandTalk integrates speech recognition and natural language understanding to allow the user to create control and force measures, to assign missions and forces, to modify missions during execution, and to control the simulation display. The current demonstration system allows setup and control of small simulated battlefield exercises by voice.

Command and Control

VALAD is a state-of-the-art interactive spoken language understanding system that adds voice activation to the Logistics Anchor Desk (LAD), an automated decision support environment that integrates relevant technologies, systems, and models into a logistical planning and execution system. The voice activation application allows military experts to perform a wide range of logistics operations quickly and naturally by adding speech understanding to LAD's existing graphical user interface. One result of this capability is increased speed in creating plans, and making it possible to develop them more effectively. The problem: Mouse-and-menu interfaces take excessive pointing, clicking, dragging, and typing strokes to get the job done. It can be difficult and time consuming to select the most appropriate items from the logistics database; for example, finding out the stock level of Avengers at the missile maintenance support of the air defence battalions. Information from multiple sources can be hard to find and tedious to use. The solution: Spoken language interfaces allow logistics planners to ask for the information they need in their own words. Phrases can refer to data from multiple sources (such as "What's the stock level of Avengers at the missile maintenance support of the air defence units?") or can specify multiple actions (such as "Show the air defence battalions in a 50 mile region around Pusan"). Speech is fast, natural and easy. With VALAD, military logistics experts can:

1 Use a large technical vocabulary comprised of thousands of words.
2 Use natural, continuous speech instead of isolated words.
3 Use the speech interface immediately, with minimal training.
4 View examples of how words and phrases can be used.

5 Save complex interactions in a "hotlist" for re-use later.
6 Get information more quickly than by using a mouse.
7 Get more kinds of data, faster.
8 Respond quickly to changing situations and requirements.
9 Get answers to the questions that really matter: "Where is...", "What is...", "When will...", etc.

Voice input extends the capabilities of the Logistics Anchor Desk. A logistics planner can short-cut lengthy sequences of operations by speaking commands and questions, and can quickly access more information than is readily accessible in the basic LAD system. The spoken language capabilities of the interface are smoothly integrated with existing mouse-and-keyboard modes of input, extending rather than limiting the options available to the user.
VALAD gives LAD ears, to hear questions and commands. The system is being developed by BBN under the sponsorship of ARPA. It demonstrates the applicability of advanced speech recognition and language understanding technology to realistic logistics tasks. By enabling users to specify information in the logistics database in natural English, the system supports the military planning process and enhances the decision-support environment of the LAD.
VALAD uses the HARK™ Recognizer, a commercial-off-the-shelf speech recognition system, to recognize spoken commands and questions. The recognizer is speaker-independent, handles continuous speech with a large vocabulary, and does not require any special hardware. It is coupled to a language understanding system that is under development at BBN to translate the recognized speech into database retrievals and commands for the LAD interface.

5.7 Voice Technology in Space: an Application in the Waiting

"Computer, run program!"
Captain James T. Kirk
USS Enterprise

For nearly three decades, the science-fiction television program Star Trek has delighted several million fans by featuring a myriad of invented technologies, ranging from teleportation to warp speed. Yet among the many illusory advances introduced in this TV program, one of the most realistic for our time and age has been the extensive use by crew members of computer voice applications such as voice control, machine translation, speaker recognition, spoken data entry, expert system interaction and more. Though such advances have yet to be perfected, speech is often argued as being the ultimate medium for human-machine interaction and its capabilities are increasingly being used in commercial applications to ease or expedite many tasks (Das and Nadas 1992). It has been long recognized in the aerospace community that some form of voice control and automatic speech recognition could be a useful addition to the collection of human-machine interfaces available today.

Unfortunately, despite the benefits it could bring to on-orbit operations, natural speech has yet to become widely used as an I/O modality in space. But the promises are real and it is probably just a matter of time until voice technologies take the inside of spacecraft by storm.

The workplace: Space

35 years ago, a young Russian Captain, *Yuri Gagarin,* was shot into orbit aboard a tiny capsule. This was the first of an incredible series of human accomplishments in space. Since then, a dozen men have walked on the moon, probes have been sent to most planets of our solar system, and a permanently habited space station orbits the Earth. The knowledge and technology acquired

through the space endeavour have been profitably put to use in world-wide communications, remote sensing and for general social and scientific advancement. Yet, the environmental conditions that prevail in space remain challenging for both humans and equipment.

Perhaps the most striking environmental feature of orbital flight is the cancellation of Earth's normal gravitational force. Weightlessness not only generates significant levels of stress in the human organism (Nicogossian et al. 1994), but transforms the entire operational conditions. In fact, micro gravity (weightlessness) destabilizes the astronauts' vestibular system and produces a broad range of behavioural consequences. Micro gravity conditions influence the interface configuration of aerospace systems and impact the design of the simplest systems. For instance, to perform any work in space, astronauts must hold themselves down one way or another, using straps and foot restraints, or simply grabbing hand holders to maintain their position.

The need to design fault-free systems and to sustain productive human operation during space flight has presented many unique challenges to aerospace engineers. More so than ever, the current generation of spacecraft requires crew members to interact with several different complex systems. Astronauts depend on the system interface for all aspects of space life including the control of the onboard environment and life support system, the conduct of experiments, the communication among the crew and with the ground, and the execution of emergency procedures. Because most systems are separately designed, the associated crew interfaces display a general lack of commonality and have to be learned separately. A typical crew interface requires specific operation procedures which are extensively detailed in checklists and repeatedly rehearsed during training. The success of the mission's objectives, as well as the safety of the spacecraft, then greatly reside in the ability of the crew to master each different system independently.

In this context, the importance of effective human computer interaction cannot be underestimated.

Applying Voice in Space

Interest in voice technology for space applications is not new. NASA and other space agencies have been pursuing applications of voice recognition and synthesis for both spacecraft and ground operations. Several testbeds have incorporated voice into their commanding scheme, but only a few experiments have been performed in operational environments (Payette 1994). In 1990, digital recordings of an astronaut's voice were performed on the ground before, during, and after a mission. A selected vocabulary was used and templates were made. After analysis, significant acoustic differences were noted, but no conclusions were drawn as to whether micro gravity was the cause of these changes in voice production, since the discrepancy was mostly blamed on a substantial difference between recording environments. A similar experiment was repeated in 1993 using a commercially available voice processing software package (Voice Navigator), with inconclusive outcome (Morris et al. 1993).

To date, only one formal speech recognition experiment has been performed aboard a spacecraft. In October 1990, NASA flew a Voice Command System (VCS) that had the capability to control the closed-circuit television cameras and monitors of the Space Shuttle (Salazar 1991). The system operated camera selection and camera functions such as pan, tilt, focus, iris and zoom. The system was speaker dependent with templates of the voice of two astronaut-operators previously made on the ground. The recognizer had limited continuous recognition and syntactic capabilities. The system was re-trained during flight to increase recognition accuracy, but the experiment showed the operational effectiveness of controlling a spacecraft subsystem using voice input. Analysis of the data showed little variation between the micro gravity and ground-based templates of the operators' voices, and astronauts stated that voice control was a useful, easy to learn tool for performing secondary tasks on the Space Shuttle. Additional data is deemed necessary to further

investigate of micro gravity on speech production and recognition. The astronaut office at NASA also states that further in-flight evaluation of the system's reliability must be performed. An upgraded version of the VCS will be flying on Space Shuttle mission STS-78 in June 1996.

The concept of a more integrated, more flexible human-oriented interaction is clearly pertinent in space application. Astronauts are functional components of space systems, not only as operators and controllers, but as contributors to the overall performance of the system. To meet the increasingly demanding mission objectives, confronted with often intricate operational procedures, astronauts undergo extensive ground training and months of thorough preparation. But further challenges lie ahead, for at the turn of the century, the International Space Station Alpha will house permanent multi ethnic crews for long periods of time. Aboard the Station, the network of computers will control and monitor thousands of automated systems as well as provide an interface to the crew. The need for performance will be heightened, necessitating increased automation and expansion of the supervisory role of the crew members.

Since few, if any, external resources and development systems will be available on a permanent space platform such as Space Station, great selectivity and perspicacity will have to be exercised when designing the human computer interface. Hence, the interest in implementing new forms of adaptable, multipurpose interfaces, such as ASR and speech synthesis. However, the decision to automate certain aspects of mission operations will demand a careful consideration of the potential human-computer relationship. The decision to use speech processing for a particular set of functions will depend on many factors such as availability, appropriateness, cost, compatibility with existing systems, and more importantly, safety and efficiency.

But the potential applications are numerous:

- Automated checklist where crew members can query procedures without having to let go of their current tasks (including speech recognition and synthesis capabilities);
- Support for telerobotic and payload applications (where operators have their hands busy and their visual attention taken by more critical tasks);
- Subsystem control, such as the video camera control aboard the Space Shuttle, which is currently performed by a second crew member beside the main operator;
- Dictation and direct data entry (speech-to-text applications) to facilitate and accelerate report generation and experiment recording;
- Speaker identification application to enhance system security;
- Automatic language translation during international missions.

Technical constraints and environmental factors will impose significant implementation requirements on the use of ASR and voice technology in space. Issues to be considered range from the technical choices (isolated word versus continuous speech, single versus multiple speakers, word based versus phoneme based, intelligibility of synthesis), the training update and maintenance requirements, the magnitude of changes in voice characteristics while in micro gravity, and the effect of high ambient noise (due to the life support system) upon maintenance of highly accurate recognition.

Perhaps the most important restriction that will be imposed on recognition systems will be a very high recognition accuracy rate. In fact, it is has been demonstrated in recent crew evaluations performed at NASA that astronauts will quickly switch back to mechanical controls if latency, reliability and efficiency criteria are not met in a new system. Also, safety requirements will impose the need for a high level of feedback to the users, with interactive error correction and on-line user query functions. Finally, on the International Space Station, the diversity of languages and accents present will undoubtedly make the implementation of voice applications an even more difficult challenge to meet.

But just as people will keep going further into space, voice will eventually make its way into spacecraft. The basic technology is on the market and will only improve with time. Its application, on the other hand, remains underestimated and the potential benefits untapped.

5.8 Speech Coders 600-1200 Bps

The NATO workgroup which is responsible for narrow-band secure-voice coding (AC302/(SG-11)/WG2) has studied the suitability of the new advanced very low bit-rate coders for use in tactical networks. For this purpose, the speech intelligibility of several very low bit-rate speech coding systems was determined (developed in various NATO countries; typical bit rate below 1200 Bps). As a reference, two existing coders were included in the evaluation.
The assessment was performed in four countries: Canada, France, The Netherlands, and the USA. The tests used for this evaluation included Mean Opinion Score tests, CVC-word tests (Consonant-Vowel-Consonant) and Diagnostic Rhyme Tests (DRT). This section presents the results of the experiments performed in The Netherlands and is focused on an intelligibility test based on CVC-words. The relation between the CVC-word score, the DRT and the related qualification can be obtained from section 4.2.

The coders, with bit rates ranging from 600-1200 Bps, were evaluated together with two reference coders labelled A and B. The reference coder B is in use in existing 2400 Bps secure voice communication systems.

Coder	Bit rate (Bps)	
A	2400	reference 1
B	2400	reference 2
C	600	
D	800	
E	1200	

As speech coding systems are normally used in a noisy environment, some of the test were also performed with additional noise at the input of the coder. In the official assessment, two types of noise at two signal-to-noise ratios were used. This review is limited to one type of noise (speech noise equivalent to voice babble) at a signal-to-noise ratio of 6 dB (about the worst that this type of system can handle). The gender of the speaker was also included as a parameter of the test. In LPC based systems, the intelligibility of female speakers is normally lower than the intelligibility of male speakers. Hence, four test conditions are described here: two signal-to-noise ratios (no noise, and 6 dB) and male and female speech.

The mean CVC-word score (m %) and the standard error (se %) are given in Table 5.1.

Table 5.1. CVC-word scores (m %) for male and female speech based upon 16 speaker-listener pairs. The standard errors (se %) are also given.

	Condition							
	Male				Female			
	No noise		SNR 6 dB		No noise		SNR 6 dB	
Coders	m	se	m	se	m	se	m	se
A	65.1	3.5	33.6	1.3	57.6	2.6	27.6	2.5
B	66.9	2.7	43.0	2.4	65.5	2.4	24.8	1.6
C	48.4	2.3	19.9	1.5	47.9	2.2	17.8	1.3
D	64.0	2.2	28.7	1.5	56.4	1.2	20.7	1.7
E	66.8	2.8	36.0	1.6	54.8	2.5	21.0	1.4

The results indicate that the coders offer a lower intelligibility for female voices. Coders working at a bit rate of 1200 and 800 Bps perform similar to the older 2400 Bps reference systems in the conditions without noise. Finally, noise has a major effect on the performance. Compared with waveform coders and analogue systems, a substantial decrease of the intelligibility is obtained. However, the low bit rate allows a fair transmission under severe jamming conditions. The systems perform at such a level that operational use is foreseen, however improvement of the intelligibility is required.

Conclusion

The primary goal of this report is to describe the military applications of speech and language processing, and the corresponding available technologies. The military applications are itemized in six categories:

- Command and Control,
- Communications,
- Computers and Information Access,
- Intelligence,
- Training which also includes language training,
- Joint Forces.

For each category a description of the requirements and possible goals are given. The available technologies are subdivided in:

- Speech Processing
- Language Processing
- Interaction
- Assessment and Evaluation.

For these technologies the state-of-the-art with respect to performance and availability is discussed. For speech processing a sub-division for speech coding, speech synthesis and recognition is made. Also an overview is given of possible assessment procedures and design criteria. Finally some case studies and applications are described.

In brief the reports highlights the need of speech control for operational systems and advanced communications in a changing military environment. Reduction of personnel, increasing complexity of systems, multi-national operations require optimal human performance in which speech can be a natural means of interfacing.

The Research Study Group which performed this study hopes that it will be a useful tool for the Operational staffs, Defence Research Staffs, and potential users within procurement departments of the NATO countries.

Reference list

AGARD Lecture series No. 170 (1990). "Speech analysis and Synthesis and Man-Machine Speech Communications for Air Operations". Eight selected tutorials by members of RSG.10. AGARD Neuilly sur Seine, France.

Boy, G.A. "Integrated Human-Machine Intelligence in Space Systems". Acta Astronautica, Vol 27, pp 175-183, 1992.

Bronkhorst, A.W., Veltman, J.A., and Breda, L. van (1996). "Application of a Three-Dimensional Auditory Display in a Flight Task". Human Factors, 38(1), 23-33.

Campbell, J.P., "Testing with the YOHO CD-ROM Voice Verification Corpus", Proc. ICASSP 95, Detroit, 1995.

Das, S., and Nadas, A., "The Power of Speech". BYTE Magazine, April 1992, pp 151-160.

Gagnon, L. (1993). "A state-based noise reduction approach for non-stationary additive interference". Speech Communication 12, p 213-219. North Holland.

Gagnon, L., and Cupples, E.W. (1995). "RSG.10 Automatic Speech Recognition in Additive Noise II", NATO Technical Report AC/243(panel 3) TR/17. Research Study Group 10 on Speech Processing.

Gish, H., and Schmidt, M., "Text-Indepedent Speaker Identification", IEEE Signal Processing Magazine, October 1994, pp 18-32.

Leeuwen, D.A. van, Berg, L.G. van den, and Steeneken, H.J.M. (1995). "Human Benchmarks for speech independent large vocabulary recognition performance". Proc. Eurospeech Madrid, 1461-1464.

Morris, R.B., Whitmore, M., and Adam, S.C., "How Well Does Voice Interaction Work in Space ?" IEEE AES Systems Magazine, August 1993, pp 26-30.

Muthusamy, Y., Barnard, E., and Cole, R., "Reviewing Automatic Language Identification", IEEE Signal Magazine, October 1994, pp 33-41.

Naylor and Porter, "An Effective Speech Separation System which Requires No A priori Information," Proc.ICASSP 91, Toronto.

Nicogossian, A.L., Huntoon, C.L., and Pool, S.L., "*Space Physiology and Medicine*". Lea & Febiger, 1994, 3rd edition.

Payette, J., "Advanced Human-Computer Interface and Voice Processing Applications in Space". In Proceedings of the ARPA Speech and Natural Language Workshop, Princeton NJ, 1994.

Ricart, R., Cupples, E.W., and Fenstermacher, "Speaker Recognition in Tactical Communications", Proc. ICASSP'94, Adelaide, Australia.

Rohlicek, J.R., "Gisting Conversational Speech", Proc. ICASSP 92, San Francisco, CA, v. 2, pp 113-116, March 1992.

Salazar, G., "Voice Recognition Makes its Debut on the NASA STS-41 Mission". Speech Technology, Feb/March 1991, pp 86-92.

Spanias, A., "Speech coding: a tutorial review", Proc. IEEE, vol. 82, pp 1341-1382, 1994.

Steeneken, H.J.M., and Houtgast, T., (1980). A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. **67** (1), 318-326.

Steeneken, H.J.M. (1992). "Quality evaluation if speech processing systems", Chapter 5 in *Digital Speech Coding: Speech coding, Synthesis and Recognition*, edited by Nejat Ince, (Kluwer Norwell USA), 127-160.

Tremain, T., (1990) "The Government Standard Linear Predictive Coding Algorithm: LPC-10", Speech Technology Magazine, pp. 40-49, April 1982.

Vaseghi, S.V., and Rayner, P.J.W., "Detection and suppression of impulsive noise in speech communication systems", IEE proceedings, Vol 137, pt 1, no. 1, p 38.

Weinstein, C.J., "Demonstrations and Applications of Spoken Language Technology", Proc. ICASSP 94.

Weinstein, C.J. (1991). "Opportunities for advanced Speech Processing in Military Computer-based Systems". NATO-DRG report, also Proc. IEEE (**79**) 1626-1641.

Zissman, M.A., "Cochannel Talker Interference Suppression", M.I.T. Lincoln Laboratory Technical Report TR-895, 26 July 1991. Also ESD-TR-91-051.

Points of Contact

The Netherlands: Dr. H.J.M. Steeneken (chairman)
TNO Human Factors Research Institute
P.O. Box 23
3769 ZG Soesterberg
Phone: +31 346 356269 Fax: +31 346 353977
E-mail: steeneken@tm.tno.nl

Canada: Mr. L. Gagnon
Communications Security Est.
Dept. of National Defence
P.O. Box 9703
Ottawa, Ontario K1G 3Z4
Phone: +1 613 991 7192 Fax: +1 613 991 7323
E-mail: lgagnon@manitou.cse.dnd.ca

Belgium: Prof. dr. C.R.A. Vloeberghs
Royal Military Academy Brussels
Renaissancelaan 30
B-1000 Brussels
Phone: +32 2 737 6245 Fax: +32 2 737 6047
E-mail: claude.vloeberghs@tele.rma.ac.be

France: Dr. D. Windheiser
DRET/STRDT/G1
26 Boulevard Victor
00460 Paris Armées
Phone: +33 1 455 24639 Fax: +33 1 455 26520
E-mail: windheis@etca.fr

Germany: Mr. F.F. Leyendecker
Amt für Nachrichtenwesen der Bundeswehr Abt I
Wolfgang-Muller Strasse 18
D-53474 Bad-Neuenahr-Ahrweiler
Phone: + 49 2225 932611 Fax: +49 2225 932609
E-mail: not available

Portugal: Prof. dr. I.M. Trancoso
INESC
R. Alves Redol, 9
1000 Lisbon
Phone: +351 1 314 5843 Fax: +351 1 310 0268
E-mail: isabel.trancoso@inesc.pt

Spain: Mr. R. Martinez
Ministerio de Defensa
Av. Padre Huidobro, km 8,500
28023 Madrid
Phone: +34 1470 2463 ext 1021 Fax: +34 1 307 8020
E-mail: not available

United Kingdom: Prof. dr. R.K. Moore
Speech Research Unit
DRA Malvern
St. Andrews Road
Great Malvern WR14 3PS
Phone: +44 1 684 894091 Fax: +44 1 684 895103
E-mail: moore@signal.dra.hmg.gb

USA: Mr. E.J. Cupples
Rome Laboratory/IRAA
32 Hangar Road
Griffiss AFB, NY 13441
Phone: +1 315 330 4024 Fax: +1 315 330 2728
E-mail: cupples@rl.af.mil

List of Authors

P. Alinàt	Thomson Sintra ASM, France
M. Bates	BBN Systems and Technologies, USA
S. Bodenkamp	Amt für Auslandsfragen, Germany
A.W. Bronkhorst	TNO Human Factors Research Institute, The Netherlands
E.J. Cupples	Rome Laboratory/IRAA, USA
L. Gagnon	Communications Security Establishment, Dept. of Nat. Defence, Canada
F.F. Leyendecker	Amt für Nachrichtenwesen der Bundeswehr Abt. I, Germany
D.A. van Leeuwen	TNO Human Factors Research Institute, The Netherlands
R. Martinez	Ministerio de Defensa, Spain
R.K. Moore	Speech Research Unit DRA, UK
G. O'Leary	MIT Lincoln Laboratory, USA
J.M. Pardo	Universidad Politécnica de Madrid, Department of Electronic Engineering, Spain
J. Payette	Canadian Space Agency, Canada
E.W. Pijpers	National Aerospace Laboratory, The Netherlands
Chr. Rouchouze	Delegation Générale pour l'Armement/Direction des Constructions Navales, France
A.M. Schaafstal	TNO Human Factors Research Institute, The Netherlands
R.W. Series	Speech Research Unit DRA, UK
H.J.M. Steeneken	TNO Human Factors Research Institute, The Netherlands
A.J. South	DRA Air systems, UK
C. Swail	Institute for Aerospace Research, National Research Council, Canada
M.M. Taylor	DCIEM, Canada
I.M. Trancoso	Institutio de Engenharia de Sistemas e Computadores, Instutio Superior Técnico, Portugal
Ph. Valéry	Sextant Avionique, France
C.R.A. Vloeberghs	Royal Military Academy Brussels, Belgium
C.J. Weinstein	MIT Lincoln Laboratory, USA
D. Windheiser	Delegation Générale pour l'Armement/Direction de la Recherche et de la Technologie, France

List of Abbreviations

ADPCM	Adaptive Pulse Code Modulation
ASR	Automatic Speech Recognition
ASW	Anti Submarine Warfare
C3I	Command Control Communications Intelligence
CELP	Code Excited Linear Predictive coding
CIC	Control Information and Command
CNI	Communication, Navigation and Identification
CVSD	Continuous Variable Slope Delta modulation
DARPA	Defence Advanced Research Project Agency (USA)
DRT	Diagnostic Rhyme Test
ELRA	European Langugage Resource Agency
EW	Electronic Warfare
HCI	Human Computer Interface
HOTAS	Hands On Throttle And Stick
LAD	Logistics Anchor Desk
LDC	Linguistic Data Consortium
LPC	Linear Predictive Coding
MOS	Mean Opinion Score
MLU	Mid Life Update
NATO	North Atlantic Treaty Organization
PCM	Pulse Code Modulation
PTT	Push To Talk
ROC	Receiver Operating Curve
RSG	Research Study Group
STANAG	Standard Agreement
STI	Speech Transmission Index
STU	Secure Telephone Unit
TDM	Time Division Multiplex
VQ	Vector Quantization

DRG DOCUMENT CENTRES

NATO does not hold stocks of DRG publications for general distribution. NATO initiates distribution of all DRG documents from its Central Registry. Nations then send the documents through their national NATO registries, sub-registries, and control points. One may sometimes obtain additional copies from these registries. The DRG Document Centres listed below can supply copies of previously issued technical DRG publications upon request.

BELGIUM
EGM-JSRL
Quartier Reine Elisabeth
Rue d'Evere,
1140 Bruxelles
Tel:(02)243 3163, Fax:(02)243 3655

CANADA
Directorate of Scientific Information Services
National Defence Headquarters
MGen. George R. Pearkes Building
Ottawa, Ontario, K1A OK2
Tel:(613)992-2263, Fax:(613)996-0392

DENMARK
Forsvarets Forskningstjeneste
Ryvangs Alle 1
2100 København Ø
Tel:3927 8888 + 5660,
Fax:3543 1086

FRANCE
CEDOCAR
00460 Armées
Tel:(1)4552 4500, Fax:(1)4552 4574

GERMANY
DOKFIZBw
Friedrich-Ebert-Allee 34
5300 Bonn 1
Tel: (0228)233091, Fax:(0228)125357

GREECE
National Defence Headquarters
R+T Section (D3)
15561 Holargos, Athens
Tel: (01)64 29 008

ITALY
MOD Italy
SEGREDIFESA IV Reparto PF.RS
Via XX Settembre, 123/A
00100 Roma
Tel:(06)735 3339, Fax:(06)481 4264

THE NETHERLANDS
KMA Bibliotheek
Postbus 90154
4800 RG Breda
MPC 71 A
Tel:(076)527-4911, Fax:(076)527-4252

NORWAY
Norwegian Defence Research Establishment
Central Registry
P.O. Box 25
2007 Kjeller
Tel:(06)80 71 41 Fax:(06)80 71 15

PORTUGAL
Direcção-General de Armamento
Ministério da Defesa Nacional
Avenida da Ilha da Madeira
1499 Lisboa
Tel:(01)610001 ext.4425, Fax:(01)611970

SPAIN
DGAM
C/ Arturo Soria 289
28033 Madrid
Tel:(91)2020640, Fax (91)2028047

TURKEY
Genelkurmay,Genel Plân Prensipler
Savunma Arastırma Daire Başkanlıgı
Ankara
Tel:(4)1176100 ext.1684, Fax:(4)11763386

UNITED KINGDOM
DRIC.
Kentigern House, 65 Brown Street
Glasgow G2 8EX
Tel:(041)224 2435, Fax:(041)224 2145

UNITED STATES
DTIC
Cameron Station
Alexandria, VA 22304-6145
Tel:(202)274-7633, Fax:(202)274-5280

DEFENCE RESEARCH SECTION
NATO HEADQUARTERS
B 1110 BRUSSELS
BELGIUM
Telephone [32](2)707 4285 - Telefax [32](2)707 4103
(not a DRG Document Distribution Centre)